

**MARGARETH EVANGELISTA BOTELHO**

**CAUSAL NETWORKS, GENOMIC PREDICTION AND CANDIDATE GENES FOR  
BOAR TAINT COMPOUNDS**

Thesis presented to the Animal Science  
Graduate Program of the Universidade Federal  
de Viçosa, in partial fulfillment of the  
requirements for degree of Doctor Scientiae.

Adviser: Renata Veroneze

Co-adviser: Marcos Soares Lopes

**VIÇOSA - MINAS GERAIS  
2020**

**Ficha catalográfica preparada pela Biblioteca Central da Universidade  
Federal de Viçosa - Câmpus Viçosa**

T

B748c Botelho, Margareth Evangelista, 1990-  
2020 Causal networks, genomic prediction and candidate genes  
for boar taint compounds / Margareth Evangelista Botelho. –  
Viçosa, MG, 2020.  
90 f. : il. (algumas color.) ; 29 cm.

Texto inglês.

Orientador: Renata Veroneze.

Tese (doutorado) - Universidade Federal de Viçosa.

Inclui bibliografia.

1. Suínos - Genética. 2. Androstenona. 3. Escatol. 4. Indol.  
5. Genômica. 6. Genes. I. Universidade Federal de Viçosa.  
Departamento de Zootecnia. Programa de Pós-Graduação em  
Zootecnia. II. Título.

CDD 22. ed. 636.40821

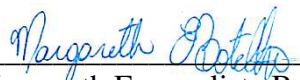
**MARGARETH EVANGELISTA BOTELHO**

**CAUSAL NETWORKS, GENOMIC PREDICTION AND CANDIDATE GENES FOR  
BOAR TAINT COMPOUNDS**

Thesis presented to the Animal Science  
Graduate Program of the Universidade Federal  
de Viçosa, in partial fulfillment of the  
requirements for degree of Doctor Scientiae.

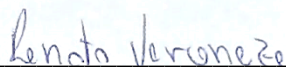
APPROVED: February 21<sup>st</sup>, 2020.

Assent:



---

Margareth Evangelista Botelho  
Author



---

Renata Veroneze  
(Adviser)

## ACKNOWLEDGMENTS

To Universidade Federal de Viçosa and the Animal Science Department, for providing me the opportunity to study in this Institution;

To *CNPq* and *Capes*, for the scholarship and financial support;

To my parents, José João and Adília; and my brothers Adenilson, Heraldo, Joelma, Gabriela, Fernando, Marta, Adelson and Everaldo for all their love, affection, encouragement and help in each moment of my life;

To my adviser, Professor Renata Veroneze, for her support, patient and enriching my knowledge;

To my co-adviser, Dr. Marcos Soares Lopes, for his support and advices;

To Professor Paulo Sávio Lopes, for his friendship and support;

To Professor Fabyano Fonseca for his patient, technical support and friendship;

To Professor Marcio de Sousa Duarte for his patient and friendship;

To all other professors, employees and students from the Animal Science Department, for making workdays a pleasant living environment;

To my friends from LABTEC and from Animal Breeding, for friendship, teaching and enjoying great moments;

To Topigs Norsvin for providing the data;

To all my friends and relatives, for praying to my health, success and happiness.

## ABSTRACT

BOTELHO, Margareth Evangelista, D. Sc., Universidade Federal de Viçosa, February, 2020. **Causal networks, genomic prediction and candidate genes for boar taint compounds.** Adviser: Renata Veroneze. Co-adviser: Marcos Soares Lopes.

The piglet non-castration may result in the boar taint appearance, which is an unpleasant taste and smell in pig meat. Boar taint is caused by the increasing of boar taint compounds (androstenone, skatole and indole) levels in adipose tissue. However, the genetic architecture and the causal relationship between levels of boar taint compounds in adipose tissue still require elucidation. In this sense, firstly, we studied the causal relationship between androstenone, skatole and indole levels in carcass adipose tissue samples and animal biopsies using structural equations models (SEM). In summary, we verified that using priori information to define the causal structure increased the model goodness-of-fit, however, the credibility intervals were also increased resulting in several unexpected null genetic correlation. We identified direct and indirect effects between boar taint compounds, mainly androstenone in biopsies affect skatole in carcass and skatole in carcass affect androstenone in carcass. Posteriorly, we evaluated the effects of SNPs weighting strategies on predictive ability and bias of genomic prediction for boar taint compounds using a single-line and multi-line populations. In general, SNP weighting strategies did not result in better predictive ability for androstenone. On the other hand, considering skatole and indole better predictive ability were archived when using weights based on gene networks. Due the slightly improvement in prediction accuracy and the increase in the number of analyses steps required, the weighting methods may not be advantageous. In addition, we verified that multi-line populations improve the prediction for androstenone, while for skatole and indole this was not observed. Finally, we performed the identification of QTLs and genes associated with boar taint compounds using a weighted single-step genome-wide association study. We used a gene network approach to improve the identification of candidate genes. In summary, we identified the *HSD17B2* gene that was previously describe as linked to boar taint appearance. New candidate genes with potential to explain boar taint phenotypes were find: *CRHBP*, *CTDSP2*, *CDK4*, *CYP27B1* e *SDR4E1*. These genes were mainly involved to biosynthesis, releasing and response to steroid hormones and intestinal absorption.

Keywords: Androstenone. Candidate gene. Genome-wide selection. Causal relationship. Indole. Skatole.

## RESUMO

BOTELHO, Margareth Evangelista, D. Sc., Universidade Federal de Viçosa, fevereiro de 2020. **Redes causais, predição genômica e genes candidatos para componentes do cheiro do varrão.** Orientadora: Renata Veroneze. Coorientador: Marcos Soares Lopes.

É desejável evitar a castração de leitões uma vez que ela pode afetar diretamente o bem-estar animal, a produção e a qualidade da carne. No entanto, a não castração pode resultar na ocorrência do cheiro de varrão, que pode ser definido como sabor e cheiro desagradáveis na carne suína, o qual prejudica a aceitação da carne pelo consumidor. O cheiro do varrão é detectado especialmente durante o cozimento, sendo causado pelo aumento do teor dos compostos lipofílicos androstenona, escatol e indol acumulados no tecido adiposo de suínos. Já se sabe que as concentrações destes compostos no tecido adiposo dependem de vários fatores como dieta, idade e genética do animal. No entanto, a arquitetura genética e a relação causal entre os níveis de compostos ligados ao cheiro do varrão na carcaça ainda requerem elucidação. Nesse sentido, na primeira parte desta tese, foram estudadas as relações causais entre os níveis de androstenona, escatol e indol em amostras de carcaça e biópsia utilizando equações estruturais. Neste estudo, verificou-se que o uso de priori para definir as estruturas causais melhora o ajuste do modelo, contudo os parâmetros estimados apresentam maior intervalo de credibilidade, resultando em correlações genéticas nulas. Foi evidenciado também a existência de efeitos diretos e indiretos entre os compostos ligados ao cheiro do varrão medidos em carcaças e biópsias, sendo consistente entre os modelos a existência do efeito da androstenona medida em biópsia no escatol medido em carcaça e o efeito do escatol em biópsias na androstenona da carcaça. Na segunda parte desta tese foram avaliados os efeitos de diferentes ponderações de SNPs para a construção da matriz de parentesco na capacidade preditiva e no viés da predição genômica dos níveis dos compostos ligados ao cheiro do varrão. Foram usadas populações baseadas em uma única linhagem ou em múltiplas linhagens, adicionalmente, foi proposto e avaliado uma nova metodologia de ponderação dos SNPs. Foi verificado que, em geral, as metodologias de ponderação do SNP avaliadas podem não melhorar a capacidade preditiva para androstenona, já para escatol e indol, a ponderação obtida a partir de redes gênicas construídas utilizando 5% dos SNPs que explicavam maior parte da variância na associação genômica melhorou levemente a capacidade preditiva em relação ao GBLUP em passo único. Devido às análises adicionais e as melhorias modestas obtidas, o uso de ponderação pode não ser vantajoso. Além disto, o uso de população composta por linhagens

múltiplas melhorou a predição para androstenona enquanto para escatol e indol isso não foi observado. Na terceira parte desta tese, foi realizada a identificação de QTLs e genes associados aos componentes do cheiro do varrão. Além disso, foi construída uma rede gênica com o intuito de verificar se os genes identificados na associação de fato estavam ligados a processos biológicos relacionados ao cheiro do varrão. Neste estudo foi identificado o gene *HSD17B2* previamente descrito como associado ao cheiro do varrão. Além disso, novos genes candidatos também foram identificados: *CRHBP*, *CTDSP2*, *CDK4*, *CYP27B1* e *SDR4E1*, sendo os principais processos biológicos em que estes genes estão envolvidos são relacionados a produção, liberação e resposta a hormônios esteroides e absorção intestinal.

Palavras-chave: Androstenona. Escatol. Gene candidato. Indol. Predição genômica. Relação causal.

## TABLE OF CONTENTS

|  |    |
|--|----|
| <b>CHAPTER 1</b> .....   | 9  |
| <b>Review</b> .....  | 9  |
| <b>1.1. Introduction</b> .....   | 9  |
| <b>1.2. Genomic studies with boar taint compounds</b> .....  | 11 |
| <b>1.3. Genomic wide selection</b> .....   | 12 |
| <b>1.4. Genetic evaluation for boar compounds</b> .....  | 14 |
| <b>1.5. Structural equations models</b> .....  | 15 |
| <b>1.6. Reference</b> .....  | 17 |
| <b>CHAPTER 2</b> .....   | 22 |
| <b>Searching for phenotypic causal networks in boar taint compounds measured in biopsies and carcasses</b> ..... | 22 |
| <b>2.1. Abstract</b> .....   | 22 |
| <b>2.2. Introduction</b> .....   | 23 |
| <b>2.3. Materials and methods</b> .....  | 25 |
| <i>Data</i> .....  | 25 |
| <i>Searching for phenotypic causal structures</i> .....  | 25 |
| <i>Fitting a Bayesian multi-trait model</i> .....  | 26 |
| <i>Building a causal structural network</i> .....  | 27 |
| <i>Fitting the structural equations models</i> .....   | 27 |
| <i>Model comparison</i> .....  | 28 |
| <b>2.4. Results</b> .....  | 29 |
| <i>Multi-trait variance components</i> .....   | 29 |
| <i>Causal structures identification</i> .....  | 31 |
| <i>Comparing models</i> .....  | 32 |
| <b>2.5. Discussion</b> .....   | 36 |
| <b>2.6. Conclusion</b> .....   | 39 |
| <b>2.7. References</b> .....   | 40 |
| <b>CHAPTER 3</b> .....   | 42 |
| <b>Applying an association weight matrix in genomic prediction of boar taint compounds</b> .....                 | 42 |
| <b>3.1. Abstract</b> .....   | 42 |
| <b>3.3. Materials and methods</b> .....  | 44 |
| <i>Data</i> .....  | 45 |
| <i>Genotypes and Quality Control</i> .....   | 45 |
| <i>Models</i> .....  | 46 |



|  |   |    |
|--|---|----|
|  | <i>Scenarios and Weighted Matrix</i> .....              | 47 |
|  | <i>Training and Validation Populations</i> .....        | 50 |
| 3.4.   | <b>Results</b> .....                                    | 50 |
|  | <i>Variance components and genetic parameters</i> ..... | 50 |
|  | <i>Evaluation of Methods</i> .....                      | 51 |
| 3.5.   | <b>Discussion</b> .....                                 | 53 |
| 3.6.   | <b>Conclusion</b> .....                                 | 56 |
| 3.7.   | <b>References</b> .....                                 | 57 |
| 3.8.   | <b>Supplementary material</b> .....                     | 61 |
| <b>CHAPTER 4</b> .....   |   | 71 |
| <b>Weighted genome-wide association study reveals new candidate genes related to steroid hormones potentially linked to boar taint</b> ..... |   | 71 |
| 4.1.   | <b>Abstract</b> .....                                   | 71 |
| 4.2.   | <b>Introduction</b> .....                               | 72 |
|  | <i>Phenotypic and genotypic data</i> .....              | 73 |
|  | <i>Statistical analyses</i> .....                       | 74 |
| 4.4.   | <b>Results</b> .....                                    | 76 |
| 4.5.   | <b>Discussion</b> .....                                 | 79 |
| 4.6.   | <b>Conclusion</b> .....                                 | 81 |
| 4.7.   | <b>References</b> .....                                 | 81 |
| 4.8.   | <b>Supplementary material</b> .....                     | 86 |
| <b>GENERAL CONCLUSION</b> .....  |   | 90 |

# CHAPTER 1

## Review

### 1.1.Introduction

Pork is the most consumed meat in many countries. The quality and quantity of meat on male pig carcass are substantially influenced by castration, in terms that non-castrated males, as an anabolic response to sexual hormones, deposit more muscle than fat in the carcass. Moreover, due to pig welfare, the non-castration has been encouraged since many countries are abolishing it from farm management (Aldal et al., 2005; Giersing et al., 2006). On the other hand, male piglet castration is practiced in production systems to avoid the boar taint appearance (Bonneau and Weiler, 2019).

Boar taint is an unpleasant taste and smell in pork from non-castrate pigs, detected especially at cooking. These smell and taste are caused by the increase of some lipophilic boar taint compounds accumulated in pig adipose tissue after sexual maturity (Aldal et al., 2005; Aluwé et al., 2011; Bridi et al., 2006; Mathur et al., 2014, 2012; Rius et al., 2005). The main boar taint compounds are androstenone ( $5\alpha$ -androst-16-ene-3-one), skatole (3-methylindole) and indoles (4-phenyl-3-butenone, p-cresol and 4-ethylphenol).

Androstenone is a steroid hormone produced and secreted by testis, which levels increase at puberty. This steroid acts as pheromones stimulating sexual behavior in the female pig and appears to be easily transferred from plasma to adipose tissue (Andresen, 2006). As other steroid hormones, androstenone is metabolized by the liver (Doran et al., 2002), however due to be a lipophilic molecule, most of this steroid is accumulated in fat cells (Haugen et al., 2012). Differences in androstenone levels in adipose tissue across pigs have been related to differences in production rate of androstenone instead of in the catabolism of the steroid (Andresen, 2006).

Skatole and indoles are produced by bacterial activity in the hind-gut, mainly through tryptophan bacterial degradation (Aldal et al., 2005; Aluwé et al., 2011; Babol et al., 2002; Claus et al., 1994). The skatole and indoles are absorbed by the intestinal mucosa into the portal vein and passes through the liver to be metabolized. Any pig may present skatole and indole production in large intestine, however, in non-castrated pigs, these products are not degraded by liver due to androstenone antagonism (Andresen, 2006). Authors have been suggested that androstenone might inhibit the liver metabolism of skatole by repressing the expression of enzymes involved in skatole and indoles metabolism (Doran et al., 2002; Zamaratskaia et al.,

2004), as consequence, non-castrated pigs will present high levels of skatole and indoles deposited in adipose tissues.

Usually, the boar taint compounds are reported as presenting great genetic variation, therefore, genetic selection against boar taint is possible (Campos et al., 2015; Drag et al., 2017; Duijvesteijn et al., 2015, 2014, 2010). Generally, the heritabilities and genetic correlations for boar taint compounds are moderate-high, moreover, genetic correlations are also positive and favorable (Campos et al., 2015; Grindflek et al., 2011; Lee et al., 2005; Mathur et al., 2014; Windig et al., 2012). Moreover, using genome-wide selection (GWS) the genetic gain could be increased due to higher accuracies for young individuals than in pedigree based genetic evaluation.

Several studies have been carried out in order to improve the knowledge on the genetic architecture and parameters for boar taint compounds (Drag et al., 2018, 2017, 2019; Duijvesteijn et al., 2015, 2014, 2010; Ramos et al., 2011; Wang and Kadarmideen, 2019). It has demonstrated different quantitative trait loci (QTL), genes and markers associated with the boar taint compounds (Drag et al., 2017, 2019; Duijvesteijn et al., 2014, 2010; Wang and Kadarmideen, 2019).

The boar taint compounds may be measured in adipose tissue through carcasses sampling, at slaughterhouses, in biopsies in live animals (Aluwé et al., 2011; Heyrman et al., 2018) and in blood plasma (Moe et al., 2009; Zamaratskaia et al., 2004). Therefore, different measurement protocols are used (Ampuero Kragten et al., 2011) resulting in different phenotypes and, as consequence, allowing to identification of different genome regions associated with the same compound in different studies.

Despite several studies considering different aspects of boar taint compounds, the genetic basis, the causal relationship between them and the modeling approach used in genetic evaluation still require elucidation in order to allow a more efficient selection. Studies about the causal relationship between androstenone, skatole and indole using structural equations (Gianola and Sorensen, 2004), for example, may improve the knowledge about direct and indirect effects among boar taint compounds. The genomic prediction may be used and strategies using SNP weighting (Wang et al., 2012) need to be evaluated in boar taint compounds. In this sense, this thesis aimed: to study the causal relationship among boar taint compounds using structural equations models; to evaluate different genomic weighted prediction strategies for boar taint compounds and; to identify QTL regions linked to boar taint compounds and to investigate the biological functions of genes under these QTLs.

## 1.2. Genomic studies with boar taint compounds

The new sequencing technologies improved the single nucleotide polymorphisms (SNPs) identification across the genome, allowing the use of genomic information under different contexts in several framework species as dairy cattle (Fang et al., 2017; Hayes et al., 2009; Sun et al., 2014), beef cattle (Aguilar et al., 2019; Piccoli et al., 2014), chicken (Fragomeni et al., 2014; Wang et al., 2014) and pigs (Campos et al., 2015; Cleveland and Hickey, 2014; Knol et al., 2016). The GWS using SNPs information, have been currently widely used in the pig breeding industry (Campos et al., 2015; Knol et al., 2016). In pig breeding, GWS studies are mainly about productive traits like growth, weight gain, carcass traits (Campos et al., 2015; Sarup et al., 2016; Tusell et al., 2019) and reproductive traits (Silva et al., 2014). Studies about GWS against boar taint compounds are still scarce.

Actually, most of genomic studies about boar taint compounds are related to understand their relationship with reproductive traits (Grindflek et al., 2011) or to find SNPs linked to quantitative trait loci (QTL) (Duijvesteijn et al., 2014; Quintanilla et al., 2003) or to identify candidates genes (Drag et al., 2017; Duijvesteijn et al., 2010). Studies have suggested that the reduction of androstenone concentration may negatively affect reproductive traits. Because of the correlation between androstenone and hormones related to fertility (estrone sulfate,  $17\beta$ -estradiol and testosterone) were very high, ranging of 0.80 to 0.95 (Grindflek et al., 2011). On the other hand, skatole and indole concentrations are strongly affected by feeding. Thus, it could be lightly easy manipulated through the diet (Visscher et al., 2018). Since skatole or indole and the main sex hormones ( $17\beta$ -estradiol, and testosterone) are poorly correlated (ranging of 0.09 to 0.28) (Grindflek et al., 2011), the selection against them may have no negative effects on this reproductive traits, although the gonadal hormones influence the prevalence of boar taint (Zamaratskaia et al., 2005).

Genes that play a role in sex steroids hormones pathways were described as candidate genes for boar taint compounds. For example, genes from cytochrome P450 (CYP) family were cited as candidate genes for boar taint (Grindflek et al., 2011; Zadinová et al., 2016). This gene family acts in metabolism and synthesis of cholesterol, steroids and other lipids (Quintanilla et al., 2003), additionally they may be indirectly involved in an oxidative phase of skatole degradation (Rowe et al., 2014; Zadinová et al., 2016). Genes from hydroxysteroid dehydrogenases of the HSD17 $\beta$  family may be related to boar taint appearance (Duijvesteijn et al., 2010; Moe et al., 2009; Rowe et al., 2014) since, in human, it is involved with synthesis of

17 beta-hydroxysteroids (Labrie et al., 1995) whose shares the same metabolic pathway to the production of androsterone in pigs. These two genes families are the most important cited, however several other genes were reported as candidate genes for boar taint (Duijvesteijn et al., 2010; Moe et al., 2009; Wang and Kadarmideen, 2019; Zadinová et al., 2016) although they did not have clear role in the boar taint appearance.

Candidate genes and markers linked to boar taint diverged throughout the research (Drag et al., 2017, 2019; Duijvesteijn et al., 2014, 2010; Wang and Kadarmideen, 2019), this probably are due differences in population structure and methodology.

### 1.3. Genomic wide selection

At last century, in animal breeding, the estimated breeding values (EBV) was obtained by Henderson's mixed model methodology (Henderson, 1984, 1963) using phenotypes and pedigree (Henderson, 1984; Henderson and Quaas, 1976). The simplest animal multi-trait model in matrix notation may be described as following:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

in which:  $\mathbf{y}$  is an observation vector;  $\boldsymbol{\beta}$  is a vector of fixed effects;  $\mathbf{a}$  is a random vector of animal genetic effects;  $\mathbf{X}$  and  $\mathbf{Z}$  are matrices of incidence;  $\mathbf{e}$  is a vector of residual random effects. The joint distribution of random effects is given by:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{e} \end{bmatrix} \sim \mathbf{N} \left\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_0 \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_0 \otimes \mathbf{I} \end{bmatrix} \right\},$$

in which:  $\mathbf{G}_0$  is genetic (co)variances matrix;  $\mathbf{A}$  is pedigree-based relationship matrix;  $\mathbf{R}_0$  is residual (co)variances matrix;  $\mathbf{I}$  is the identity matrix.

This approach uses only phenotypes and pedigree to predict EBV for the selection candidates. More recently, at the beginning of the 21<sup>st</sup> century, due the development of high throughput genotyping, thousands of SNP markers densely covering the genome were identified in several species, allowing the implementation of the GWS proposed by Meuwissen et al. (2001). Genome wide selection assumed that at least part of these SNP are in linkage disequilibrium (LD) with QTLs, therefore, these markers may be used to predict genomic estimated breeding values (GEBV). The GBLUP is a common procedure used in GWS. In this method the SNPs are used to obtained the realized genomic relationship matrix ( $\mathbf{G}$ ) that replace the pedigree based relationship matrix ( $\mathbf{A}$ ) in the Henderson's mixed model equations (VanRaden, 2008). The  $\mathbf{G}$  matrix can be calculated as following:

$$\mathbf{G} = \frac{\mathbf{MDM}'}{2 \sum pq},$$

in which:  $\mathbf{M}$  is a centered matrix constructed by subtracting  $\mathbf{P}$  from  $\mathbf{X}$ , in which  $\mathbf{X}$  is a matrix that specify which marker alleles each individual inherited and  $\mathbf{P}$  contain allele frequencies expressed as a difference from 0.5 and multiplied by 2, such that column  $i$  of  $\mathbf{P}$  is  $2(p_i - 0.5)$ ;  $\mathbf{D}$  is a identity matrix;  $p$  and  $q$  are the SNP allele frequencies in each locus.

A drawback for this method is that it requires all animals to be genotyped, thus the genetic evaluation, which usually includes animals with and without genotypes, must be carried out through a multiple-step procedure (Hayes et al., 2009; VanRaden, 2008)(Misztal et al., 2009). Alternatively, Misztal et al. (2009) proposed the single-step GBLUP (ssGBLUP), in which the matrix is replaced by an  $\mathbf{H}$  matrix in the mixed model equations. The  $\mathbf{H}$  matrix was elaborated considering simultaneously animals with and without genotype information. Summarizing,  $\mathbf{H}$  matrix combining genomic ( $\mathbf{G}$ ) and pedigree based ( $\mathbf{A}$ ) relationship matrices.

Considering  $\mathbf{A}$  matrix built as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

wherein the subsection  $\mathbf{A}_{11}$  is composed by the relationship among animals that have only pedigree information,  $\mathbf{A}_{22}$ , is composed by relationship among animals that have genotype information,  $\mathbf{A}_{12}$ , and  $\mathbf{A}_{21}$  are composed by relationship among animals considered in matrices  $\mathbf{A}_{11}$  and  $\mathbf{A}_{22}$  (Aguilar et al., 2010).

The inverse of  $\mathbf{H}$  matrix ( $\mathbf{H}^{-1}$ ) is given by:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

in which:  $\mathbf{A}^{-1}$  is the inverse of  $\mathbf{A}$ ;  $\mathbf{G}^{-1}$  is the inverse of  $\mathbf{G}$ ;  $\mathbf{A}_{22}^{-1}$  is the inverse of  $\mathbf{A}_{22}$ .

Recently, an approach to attribute different SNPs weights to build the relationship matrix ( $\mathbf{G}$ ) have been evaluated (Wang et al., 2014; Zhang et al., 2016). This SNP weighting may increase the genomic prediction accuracy (Marques et al., 2018; Veroneze et al., 2016; Wang et al., 2014), however, this procedure may have different impacts on results from different traits and under divergent populations structures (Lourenco et al., 2017).

#### **1.4. Genetic evaluation for boar compounds**

There are few studies about genetic evaluation for boar taint compounds (Campos et al., 2015; Luki et al., 2015; Sellier et al., 2000). Generally, these studies evaluate mainly androstenone and skatole together to sexual or carcass traits.

Sellier et al. (2010), studied the responses to index selection for androstenone level in fat and sexual (maturity measured as bulbo-urethral gland (BUG) size) in young boar from an experimental population across four generations. The experimental design included a control and a select line having a Large White–Landrace genetic background. The index combining the average thickness of right and left BUG, the log of androstenone level (LAND) from back fat sample taken in the neck region of pigs at 117.6 kg body weight ( $I = 100 + 4 * TBUG - 63 * LAND$ ). The pattern of direct responses to antagonistic selection consisted of no response in fat LAND level and a significant positive genetic trend in BUG development. The authors observed that the selection index used throughout the experiment partly explains the pattern of direct responses to selection. They found evidences for a significant genetic relationship between age at sexual maturation of boars and gilts. This research also demonstrated that the joint selection for sexual maturity and androstenone might have unfavorable effect on sexual development.

Despite the of the possible association suggested by Sellier et al. (2010), research has suggested that low-androstenone haplotype on pig does not unfavorably affect production and reproduction traits (Hidalgo et al., 2014). These authors investigated a single nucleotide polymorphism marker distinguishing the Asian from European pig haplotypes. They found a favorable effect at least one sow line on number of teats and number of spermatozoa per ejaculation for the low-androstenone haplotype. Therefore, the unfavorable association between reproductive characteristics and androstenone levels may be considered inconsistent.

Considering the GWS approaches, Luki et al. (2015) evaluated the prediction for skatole and androstenone through GBLUP and Bayesian regression methods using phenotypes from 1000 pigs genotyped for 62,153 SNPs. These authors showed that the Bayesian approach present slightly higher predictions accuracies. For androstenone, the GBLUP presented accuracy close to the most accurate method, however for skatole, the Bayesian method provided significantly higher accuracy and should be preferable. In addition, they also showed that the whole-genome evaluation methods gave greater accuracy than using only the detected QTL in the model.

Campos et al. (2015) compared the Ridge Regression BLUP (RR-BLUP) and Bayesian LASSO (BL) methods to predict the GEBV for carcass traits and concentrations of

androstenone and skatole using information from 622 boars and 2,500 SNPs. The best GEBV accuracies for most of the traits were achieved by BL. However, the small database size may have influenced this result.

### 1.5. Structural equations models

The application of structural equations models (SEM) in quantitative genetic was proposed by Gianola e Sorensen (2004). The SEM tests the plausibility of a theoretical model about causal relationships between variables that supports the studied phenomenon. In this way, a trait may be described as function of others studied traits, allowing the elaboration of a functional network between them.

Differently from multi-trait models, the SEM use several types of models to describe the relationship between the observed variables. The idea is to provide a quantitative test of a theoretical model (Schumacker and Lomax, 2004) to evaluate and estimate the direct, indirect and total effects that one variable exert over another (Codes, 2005). SEM allow understanding how the values of some traits are affected by (and not only associated with) the values of other traits (Valente and Rosa 2013).

When many traits are evaluated, like in breeding programs, structural equations may be used as a differentiated approach of multi-trait models, making possible to combine information in cause and effect in order to simplify complex relationships. Generalizing, the structural equations can be represented as described by Valente et al. (2010):

$$y_j = f(y_{pj}, e_j)$$

in which:  $y_j$  is a dependent variable;  $y_{pj}$  are the parent variables of  $y_j$ , in other words, variables that influence  $y_j$ ;  $e_j$  is a random residual term associated with  $y_j$ .

The variables associations are represented by arrows in a structural causal diagram (Pearl, 2003). In summary, each diagram variable is represented by a vertex and are connected to others by edges that indicate a causal association. These edges may be directed edges when the edges have arrows in one extremity ( $\rightarrow$  or  $\leftarrow$ ); symmetrical direct edges, when the edges have arrows in both extremities ( $\leftrightarrow$ ); and undirected edges when the edges have no arrows ( $-$ ). Two vertexes connected by one edge are called adjacent (Valente et al., 2010). When more than two vertexes are in sequence a path is formed: if the arrow point in the same direction (ex.:  $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ ) the path is called directed path, on otherwise (ex.:  $A \rightarrow B \rightarrow C \leftarrow D \rightarrow E$ ) is undirected path. In the example  $A \rightarrow B \rightarrow C \leftarrow D \leftarrow E$ , the variable “C” is called collider.



The “parent” variables (which affect other traits) should be defined a priori and a way of identifying these variables is using the ICA (Inductive Causation algorithm) (Pearl, 2003). This algorithm considers the partial correlations to establish causality relations giving as output a causal diagram (Figure 1). Briefly, diagrams are generated in three steps. In the first step, partial correlations will be used to identify the non-directed connection between two adjacent variables and diagrams are generated in which variables are only linked (eg,  $y_1$ - $y_2$ ). In the second step, colliders variables are identified and edges previous obtained are oriented. In the third step, when possible, non-directed edges will be oriented to new unshielded colliders. This step is only necessary when the diagram obtained in step 2 shows non-oriented edges.

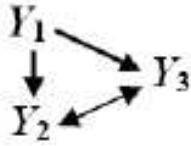


Figure 1: Representation of a structural causal diagram.

In mixed model context, Gianola and Sorensen (2004) described a two traits structural equation system as follow:

$$y_{i1} = \lambda_{12}y_{i2} + x'_{i1}\beta_1 + a_{i1} + e_{i1}$$

$$y_{i2} = \lambda_{21}y_{i1} + x'_{i2}\beta_2 + a_{i2} + e_{i2}$$

in which  $\beta_1$  e  $\beta_2$  are fixed effects vectors for trait 1 and 2, with incidence vectors  $x'_{i1}$  and  $x'_{i2}$ , respectively;  $a_{i1}$  and  $a_{i2}$  are additive genetic effects;  $e_{i1}$  and  $e_{i2}$  are random residual;  $\lambda_{12}$  is the change on  $y_{i1}$  as function of  $y_{i2}$  and  $\lambda_{21}$  is the change on  $y_{i2}$  as function of  $y_{i1}$ . This model can be written in matrix notation according to presented by Valente et al. (2010):

$$\mathbf{y} = (\mathbf{A} \otimes \mathbf{I})\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}.$$

With distribution:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{e} \end{bmatrix} \sim \mathbf{N} \left\{ \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_0 \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi}_0 \otimes \mathbf{I} \end{bmatrix} \right\}$$

in which  $\mathbf{y}$  is an observation vector containing all  $t$  traits;  $\mathbf{A}$  is a structural coefficient matrix, a square matrix of order  $t$ ;  $\boldsymbol{\beta}$  is a fixed effect vector;  $\mathbf{X}$  and  $\mathbf{Z}$  are incidence matrices;  $\mathbf{a}$  is additive genetic values vector;  $\mathbf{e}$  is a random residual vector;  $\mathbf{G}_0$  is additive genetic (co)variance matrix and;  $\boldsymbol{\Psi}_0$  is a residual diagonal matrix.

SEM can be reduced by solving it for the terms containing  $y_i$  in the right-hand side. This can be performed by transforming the model to  $(\mathbf{I} - \mathbf{\Lambda} \otimes \mathbf{I})\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}$ . Thus, we can infer that structural equations model described above is an extension of multi-trait models with the advantage of allowing to build functional networks among traits. Using this approach, it would be possible describe how each boar taint compound affect the others adding information about causality and direction effects that cannot be achieved in conventional multi-trait models.

## 1.6. Reference

- Aguilar, I., A. Legarra, F. Cardoso, Y. Masuda, D. Lourenco, et al. 2019. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. *Genet Sel Evol* 51(1): 1–8. doi: 10.1186/s12711-019-0469-3.
- Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, et al. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J Dairy Sci* 93(2): 743–752. doi: 10.3168/jds.2009-2730.
- Aldal, I., Ø. Andresen, A.K. Egeli, J. Haugen, A. Grødum, et al. 2005. Levels of androstenone and skatole and the occurrence of boar taint in fat from young boars. *Livest Prod Sci* 95(1–2): 121–129. doi: 10.1016/j.livprodsci.2004.12.010.
- Aluwé, M., S. Millet, K.M. Bekaert, F.A.M.M. Tuytens, L. Vanhaecke, et al. 2011. Influence of breed and slaughter weight on boar taint prevalence in entire male pigs. *Animal* 5(8): 1283–1289. doi: 10.1017/S1751731111000164.
- Ampuero Kragten, S., B. Verkuylen, H. Dahlmans, M. Hortos, J.A. Garcia-Regueiro, et al. 2011. Inter-laboratory comparison of methods to measure androstenone in pork fat. *Animal* 5(10): 1634–1642. doi: 10.1017/S1751731111000553.
- Andresen, Ø. 2006. Boar taint related compounds : Androstenone / skatole / other substances. *Acta Vet Scand* 48: 1–4. doi: 10.1186/1751-0147-48-S1-S5.
- Babol, J., E.J. Squires, and E.A. Gullett. 2002. Factors affecting the level of boar taint in entire male pigs as assessed by consumer sensory panel. *Meat Sci* 61: 33–40. doi: 10.1016/s0309-1740(01)00159-0.
- Bonneau, M., and U. Weiler. 2019. Pros and cons of alternatives to piglet castration: Welfare, boar taint, and other meat quality traits. *Animals* 9(11): 1–12. doi: 10.3390/ani9110884.
- Bridi, A.M., A.R. De Oliveira, N. Aparecida, N. Fonseca, L.L. Coutinho, et al. 2006. Efeito do genótipo halotano , da ractopamina e do sexo do animal na qualidade da carne suína. *Rev Bras Zootec* 35(5): 2027–2033.
- Campos, C.F. de, M.S. Lopes, F.F. e Silva, R. Veroneze, E.F. Knol, et al. 2015. Genomic selection for boar taint compounds and carcass traits in a commercial pig population. *Livest Sci* 174: 10–17. doi: 10.1016/j.livsci.2015.01.018.
- Claus, R., U. Weiler, and A. Herzog. 1994. Physiological Aspects of Androstenone and Skatole Formation in the Boar A Review with Experimental Data. *Meat Sci* 38: 289–305. doi: 10.1016/0309-1740(94)90118-X.

- Cleveland, M.A., and J.M. Hickey. 2014. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. : 3583–3592. doi: 10.2527/jas2013-6270.
- Codes, A.L.M. 2005. Modelagem de Equações Estruturais: um método para a análise de fenômenos complexos. *Cad CRH* 18(45): 471–484.
- Doran, E., F.W. Whittington, J.D. Wood, and J.D. Mcgivan. 2002. Cytochrome P450IIE1 ( CYP2E1 ) is induced by skatole and this induction is blocked by androstenone in isolated pig hepatocytes. *Chem Biol Interact* 140: 81–92. doi: 10.1016/S0009-2797(02)00015-7.
- Drag, M., M.B. Hansen, and H.N. Kadarmideen. 2018. Systems genomics study reveals expression quantitative trait loci, regulator genes and pathways associated with boar taint in pigs. *PLoS One* 13(2): 1–30. doi: 10.1371/journal.pone.0192673.
- Drag, M.H., L.J.A. Kogelman, H. Maribo, L. Meinert, P.D. Thomsen, et al. 2019. Characterization of eQTLs associated with androstenone by RNA sequencing in porcine testis. *Physiol Genomics* 51(10): 488–499. doi: 10.1152/physiolgenomics.00125.2018.
- Drag, M., R. Skinkytė-Juskienė, D.N. Do, L.J.A. Kogelman, and H.N. Kadarmideen. 2017. Differential expression and co-expression gene networks reveal candidate biomarkers of boar taint in non-castrated pigs. *Sci Rep* 7(1): 1–18. doi: 10.1038/s41598-017-11928-0.
- Duijvesteijn, N., E.F. Knol, and P. Bijma. 2014. Boar taint in entire male pigs : A genomewide association study for direct and indirect genetic effects on androstenone. *J Anim Sci* 92: 4319–4328. doi: 10.2527/jas2014-7863.
- Duijvesteijn, N., E.F. Knol, and P. Bijma. 2015. Direct and associative effects for androstenone and genetic correlations with backfat and growth in entire male pigs. *J Anim Sci*: 2465–2475. doi: 10.2527/jas2011-4625.
- Duijvesteijn, N., E.F. Knol, J.W.M. Merks, R.P.M.A. Crooijmans, M.A.M. Groenen, et al. 2010. A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. *BMC Genet* 11(42): 1–11. doi: 10.1186/1471-2156-11-42.
- Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu, et al. 2017. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet Sel Evol* 49(1): 44. doi: 10.1186/s12711-017-0319-0.
- Fragomeni, B. de O., I. Misztal, D.L. Lourenco, I. Aguilar, R. Okimoto, et al. 2014. Changes in variance explained by top SNP windows over generations for three traits in broiler chicken. *Front Genet* 5(OCT): 1–7. doi: 10.3389/fgene.2014.00332.
- Gianola, D., and D. Sorensen. 2004. Quantitative Genetic Models for Describing Simultaneous and Recursive Relationships Between Phenotypes. *Genetics* 167(3): 1407–1424. doi: 10.1534/genetics.103.025734.
- Giersing, M., J. Ladewig, and B. Forkman. 2006. Animal Welfare Aspects of Preventing Boar Taint. *Acta Vet Scand* 48(Suppl 1): S3. doi: 10.1186/1751-0147-48-S1-S3.
- Grindflek, E., T.H.E. Meuwissen, T. Aasmundstad, H. Hamland, M.H.S. Hansen, et al. 2011. Revealing genetic relationships between compounds affecting boar taint and reproduction in pigs. *J Anim Sci* 89(3): 680–692. doi: 10.2527/jas.2010-3290.
- Haugen, J.E., C. Brunius, and G. Zamaratskaia. 2012. Review of analytical methods to measure boar taint compounds in porcine adipose tissue: The need for harmonised

- methods. *Meat Sci* 90(1): 9–19. doi: 10.1016/j.meatsci.2011.07.005.
- Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J Dairy Sci* 92(2): 433–443. doi: 10.3168/jds.2008-1646.
- Henderson, C.R. 1963. Selection Index and Expected Genetic Advance. *Stat Genet Plant Breed* 982: 141–163.
- Henderson, C.R. 1984. *Applications of Linear Models in Animal Breeding Models*.
- Henderson, C.R., and R.L. Quaas. 1976. Multiple trait evaluation using relatives' records. *J Anim Sci* 43(6): 1188–1197.
- Heyrman, E., S. Millet, F.A.M. Tuytens, B. Ampe, S. Janssens, et al. 2018. On farm intervention studies on reduction of boar taint prevalence: Feeding strategies, presence of gilts and time in lairage. *Res Vet Sci* 118(February): 508–516. doi: 10.1016/j.rvsc.2018.05.008.
- Hidalgo, A.M., J.W.M. Bastiaansen, B. Harlizius, E.F. Knol, M.S. Lopes, et al. 2014. Asian low-androstenone haplotype on pig chromosome 6 does not unfavorably affect production and reproduction traits. *Anim Genet* 45(6): 874–877. doi: 10.1111/age.12226.
- Knol, E.F., B. Nielsen, and P.W. Knap. 2016. Genomic selection in commercial pig breeding. *Anim Front* 6(1): 15–22. doi: 10.2527/af.2016-0003.
- Labrie, Y., F. Durocher, Y. Lachance, C. Turgeon, J. Simard, et al. 1995. The Human Type II 17 $\beta$ -Hydroxysteroid Dehydrogenase Gene Encodes Two Alternatively Spliced mRNA Species. *DNA Cell Biol* 14(10): 849–861. doi: 10.1089/dna.1995.14.849.
- Lee, G.J., A.L. Archibald, A.S. Law, S. Lloyd, J. Wood, et al. 2005. Detection of quantitative trait loci for androstenone, skatole and boar taint in a cross between Large White and Meishan pigs. *Anim Genet* 36(1): 14–22. doi: 10.1111/j.1365-2052.2004.01214.x.
- Lourenco, D.A.L., B.O. Fragomeni, H.L. Bradford, I.R. Menezes, J.B.S. Ferraz, et al. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. *J Anim Breed Genet* 134(6): 463–471. doi: 10.1111/jbg.12288.
- Luki, B., S.J. Rowe, D.J. De Koning, I. Velander, and C.S. Haley. 2015. Efficiency of genomic prediction for boar taint reduction in Danish Landrace pigs. : 607–616. doi: 10.1111/age.12369.
- Marques, D.B.D., J.W.M. Bastiaansen, M.L.W.J. Broekhuijse, M.S. Lopes, E.F. Knol, et al. 2018. Weighted single-step GWAS and gene network analysis reveal new candidate genes for semen traits in pigs. *Genet Sel Evol* 50(1): 1–14. doi: 10.1186/s12711-018-0412-z.
- Mathur, P.K., J. ten Napel, R.E. Crump, H.A. Mulder, and E.F. Knol. 2014. Genetic relationship between boar taint compounds, human nose scores, and reproduction traits in pigs. *J Anim Breed Genet* 91(9): 4080–4089. doi: <https://doi.org/10.2527/jas.2013-6478>.
- Mathur, P.K., J. ten Napel, S. Bloemhof, L. Heres, E.F. Knol, et al. 2012. A human nose scoring system for boar taint and its relationship with androstenone and skatole. *Meat Sci* 91(4): 414–422. doi: 10.1016/j.meatsci.2012.02.025.
- Meuwissen, T.H.E., B. J. Hayes, and M.E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.

- Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J Dairy Sci* 92(9): 4648–4655. doi: 10.3168/jds.2009-2064.
- Moe, M., S. Lien, T. Aasmundstad, T.H. Meuwissen, M.H. Hansen, et al. 2009. Association between SNPs within candidate genes and compounds related to boar taint and reproduction. *BMC Genet* 10(1): 32. doi: 10.1186/1471-2156-10-32.
- Pearl, J. 2003. *Causality: Models, Reasoning, and Inference*. Econometric Theory. p. 675–685
- Piccoli, M.L., J. Braccini, F.F. Cardoso, M. Sargolzaei, S.G. Larmer, et al. 2014. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. *BMC Genet* 15(1): 157. doi: 10.1186/s12863-014-0157-9.
- Quintanilla, R., O. Demeure, J.P. Bidanel, D. Milan, N. Iannuccelli, et al. 2003. Detection of quantitative trait loci for fat androstenone levels in pigs. *J Anim Sci* 81(2): 385–394. doi: 10.2527/2003.812385x.
- Ramos, A.M., N. Duijvesteijn, E.F. Knol, J.W.M. Merks, H. Bovenhuis, et al. 2011. The distal end of porcine chromosome 6p is involved in the regulation of skatole levels in boars. *BMC Genet* 12. doi: 10.1186/1471-2156-12-35.
- Reverter, A., M.R.S. Fortes, B.D. Valente, and G.J. de M. Rosa. 2013. *Genome-Wide Association Studies and Genomic Prediction* (C. Gondro, J. van der Werf, and B. Hayes, editors). Humana Press, Totowa, NJ.
- Rius, M.A., M. Hortós, and J.A. García-Regueiro. 2005. Influence of volatile compounds on the development of off-flavours in pig back fat samples classified with boar taint by a test panel. *Meat Sci* 71(4): 595–602. doi: 10.1016/j.meatsci.2005.03.014.
- Rowe, S.J., B. Karacaören, D.-J. de Koning, B. Lukic, N. Hastings-Clark, et al. 2014. Analysis of the genetics of boar taint reveals both single SNPs and regional effects. *BMC Genomics* 15(1): 424. doi: 10.1186/1471-2164-15-424.
- Sarup, P., J. Jensen, T. Ostensen, M. Henryon, and P. Sørensen. 2016. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet* 17(1): 11. doi: 10.1186/s12863-015-0322-9.
- Schumacker, R.E., and R.G. Lomax. 2004. *Structural Equation Modeling Third Edition*. 3o.
- Sellier, P., P. Le Roy, M.N. Fouilloux, J. Gruand, and M. Bonneau. 2000. Responses to restricted index selection and genetic parameters for fat androstenone level and sexual maturity status of young boars. *Livest Prod Sci* 63(3): 265–274. doi: 10.1016/S0301-6226(99)00127-X.
- Silva, F.F., H.A. Mulder, E.F. Knol, M.S. Lopes, S.E.F. Guimarães, et al. 2014. Sire evaluation for total number born in pigs using a genomic reaction norms approach. *J Anim Sci* 92(9): 3825–3834. doi: 10.2527/jas.2013-6486.
- Sun, C., P.M. VanRaden, J.B. Cole, and J.R. O’Connell. 2014. Improvement of Prediction Ability for Genomic Selection of Dairy Cattle by Including Dominance Effects (W. Barendse, editor). *PLoS One* 9(8): e103934. doi: 10.1371/journal.pone.0103934.
- Tusell, L., H. Gilbert, Z.G. Vitezica, M.J. Mercat, A. Legarra, et al. 2019. Dissecting total genetic variance into additive and dominance components of purebred and crossbred pig traits. *Animal*: 1–11. doi: 10.1017/S1751731119001046.

- Valente, B.D., G.J.M. Rosa, G. de Los Campos, D. Gianola, and M.A. Silva. 2010. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics* 185(2): 633–644. doi: 10.1534/genetics.109.112979.
- VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. *J Dairy Sci* 91(11): 4414–4423. doi: 10.3168/jds.2007-0980.
- Veroneze, R., P.S. Lopes, M.S. Lopes, A.M. Hidalgo, S.E.F. Guimarães, et al. 2016. Accounting for genetic architecture in single- and multipopulation genomic prediction using weights from genomewide association studies in pigs. *J Anim Breed Genet* 133(3): 187–196. doi: 10.1111/jbg.12202.
- Visscher, C., A. Kruse, S. Sander, C. Keller, J. Mischok, et al. 2018. Dietary approaches reducing boar taint-Importance of *Lawsonia intracellularis* colonisation for interpreting results. *J Anim Physiol Anim Nutr (Berl)* 102(October 2017): 3–15. doi: 10.1111/jpn.12860.
- Wang, X., and H.N. Kadarmideen. 2019. Genome-wide DNA methylation analysis using next-generation sequencing to reveal candidate genes responsible for boar taint in pigs. *Anim Genet*. doi: 10.1111/age.12842.
- Wang, H., I. Misztal, I. Aguilar, A. Legarra, R.L. Fernando, et al. 2014. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Front Genet* 5(MAY): 1–10. doi: 10.3389/fgene.2014.00134.
- Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet Res (Camb)* 94(2): 73–83. doi: 10.1017/S0016672312000274.
- Windig, J.J., H.A. Mulder, J. ten Napel, E.F. Knol, P.K. Mathur, et al. 2012. Genetic parameters for androstenone, skatole, indole, and human nose scores as measures of boar taint and their relationship with finishing traits. *J Anim Sci* 90(7): 2120–2129. doi: 10.2527/jas.2011-4700.
- Zadinová, K., R. Stupka, A. Stratil, J. Čítek, K. Vehovský, et al. 2016. Boar taint – the effects of selected candidate genes associated with androstenone and skatole levels – a review. *Anim Sci Pap Reports* 34(2): 107–128.
- Zamaratskaia, G., J. Babol, H. Andersson, and K. Lundström. 2004. Plasma skatole and androstenone levels in entire male pigs and relationship between boar taint compounds, sex steroids and thyroxine at various ages. *Livest Prod Sci* 87(2–3): 91–98. doi: 10.1016/j.livprodsci.2003.09.022.
- Zamaratskaia, G., L. Rydhmer, G. Chen, A. Madej, H. Andersson, et al. 2005. Boar Taint is Related to Endocrine and Anatomical Changes at Puberty but not to Aggressive Behaviour in Entire Male Pigs. *Reprod Domest Anim* 40(6): 500–506. doi: 10.1111/j.1439-0531.2005.00613.x.
- Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. *Front Genet* 7(AUG): 1–14. doi: 10.3389/fgene.2016.00151.

## CHAPTER 2

### Searching for phenotypic causal networks in boar taint compounds measured in biopsies and carcasses

#### 2.1. Abstract

Boar taint compounds (androstenone, skatole and indole) are usually measured in pig carcass, after slaughter. Alternatively, subcutaneous adipose tissue biopsies could be used to evaluate the content of these compounds in live animals. However, it is necessary to understand how the values observed in live animals reflect the values measured in carcass. This understanding may be acquired applying an Inductive Causation algorithm (ICA) to study the causality relationship among this set of traits. The ICA allows identifying the causality among traits and describes them through a functional causal graph. This graph provides causal association information to elaborate structural equations model (SEM), allowing the estimation of direct and indirect effects between traits by the structural coefficients. Therefore, we aimed to search for the causal relationship among boar taint compounds (androstenone, skatole and indole) measured in pig adipose tissue from carcasses and biopsies. We used information of androstenone, skatole and indole compounds measured in 3,590 adipose tissue samples from pig carcasses (AC, SC and IC, respectively) and 397 adipose tissue samples from biopsies (AB, SB and IB, respectively). We fitted a multi-trait model and SEM considering causal networks graphs obtained by ICA with or without *a priori* information. The priori considered was: biopsies are realized in live animal, before slaughter, therefore consist of previous observations exerting effects on carcass observations. The models were compared through the Deviance Information Criterion (DIC). The best DIC was obtained in a model with the causal structure consisted of  $SC \leftarrow AB \rightarrow AC \leftarrow SB$ , however, this structure was built using a priori information and the SEM fitted with this priori returned several null genetics correlations among traits well described as positive correlated. The best structure returned using only ICA was  $IB \rightarrow SC \leftarrow AB \leftarrow AC \leftarrow SB: SC \rightarrow IC$ , which was obtained with 80% to 70% high probability distribution (HPD) interval. This model returned positive genetic correlations and a small loss in the model goodness-of-fit was obtained. The use of causal structure to build the SEM improved the model goodness-of-fit compared with the multi-trait model in all the cases, indicating that the SEM was more plausible. Even using small HPD interval contents to build causal networks, we identified causal relationships among boar taint compounds in carcasses

and biopsies. The boar taint compounds measured in biopsies exert direct effects on boar taint compounds measured in carcasses and are passive to be used in breeding programs, improving the model goodness-of-fit in the prediction of these compounds.

Keywords: Biopsies. Carcasses. Causal structure. Structural equation models.

## 2.2. Introduction

Boar taint, an unpleasant taste and smell of pork, is caused by the increase of androstenone, skatole and indole (boar taint compounds) levels (Aluwé et al., 2011; Mathur et al., 2014, 2012; Rius et al., 2005), especially in pig adipose tissue. Non-castrated pigs produce androstenone in the testis, which avoids skatole degradation in the liver. Consequently, these pigs have greater androstenone and skatole depositions than barrows and gilts (Zamaratskaia and Squires, 2009). Similar to skatole, indole is synthesized from the tryptophan metabolism in the hindgut (Rius et al., 2005). Both compounds are absorbed and metabolized in the liver and the non-metabolized excess is deposited in adipose tissue.

Boar taint compounds are usually measured in the animal adipose tissue at slaughterhouses. In this sense, these traits can only be measured in the relatives of the selection candidates, which may decrease the genetic gain. Alternatively, such compounds can be measured on live animals through subcutaneous adipose tissue biopsies. Thus, it would be of valuable interest to understand how boar taint compounds measured in live animals reflect the boar taint compounds measured in carcass.

It is already known that boar taint compounds in carcass adipose tissue present great genetic variability, with heritabilities ranging from 0.46 to 0.72, 0.29 to 0.35 and 0.26 to 0.50 for androstenone, indole and skatole, respectively (Campos et al., 2015; Grindflek et al., 2011; Lee et al., 2005; Mathur et al., 2014; Windig et al., 2012). In addition, studies have reported moderate positive and favorable genetic correlations (ranging from 0.30 to 0.62) between androstenone and indole/skatole (Campos et al., 2015; Lee et al., 2005; Mathur et al., 2014; Windig et al., 2012), and stronger positive and favorable correlations between indole and skatole, ranging from 0.71 to 0.78 (Grindflek et al., 2011; Lee et al., 2005). However, to our knowledge, there are no studies reporting heritabilities for boar taint compounds measured in biopsies, and the genetic correlations between these compounds in carcasses and biopsies.

A causality relationship study can be performed to help understanding how values observed in live animals are related to the values measured in carcass. In this context, Gianola



and Sorensen (2004) proposed the application of structural equations in quantitative genetics to study the causal relationships among several traits. Structural equations model (SEM) may be used as an alternative approach to the multi-trait models (MTM) in animal breeding programs, with the advantage of using causal relationship structures to describe the most likely relationship between the observed variables. The idea is to provide a quantitative test for a theoretical model (Schumacker and Lomax, 2010) to evaluate and estimate direct, indirect and total effects that one variable exerts on another variable (Gianola and Sorensen, 2004). The SEM allows understanding how the phenotypes for some traits are affected by (and not only associated with) other traits (Valente et al., 2011) and has the advantage of allowing to build functional networks among traits.

Fitting SEM, however, requires *a priori* choice of a causal structure. One way to search for recursive causal structures is to apply the Inductive Causation algorithm (ICA) (Pearl, 2003). As output, this algorithm returns a causal relationship graph in which a vertex represents each variable and their associations are represented by arrows. The ICA is constructed based on specific assumptions that are not completely fulfilled by multiple phenotypes in genetic evaluation, since traits may present unobserved correlated genetic effects that confound the search for causal structures (Valente et al., 2011, 2010). In this sense, Valente et al. (2011) proposed a methodology based on fitting a Bayesian model and the application of the ICA (Pearl, 2003) to the joint distribution of phenotypes conditional on genetic effects. Briefly, the partial residual correlations are used to identify the non-directed connection between two variables and graphs in which traits ( $y_1$  and  $y_2$ ) are linked by edges are generated (e.g.,  $y_1$ - $y_2$ ). Then, collider variables are identified and the edges previous obtained are oriented (e.g.,  $y_1 \rightarrow y_2$ ). Finally, when possible, non-directed edges will be oriented to new unshielded colliders. The causal relationship graph obtained by IC provide the *priori* information to elaborate the SEM, allowing estimation of direct and indirect effects between traits by the structural coefficients obtained by solving the SEM (Valente et al., 2011, 2010).

Although genetic parameters have been estimated for boar taint content in carcass, the causal relationship between levels of androstenone, skatole and indole in biopsies and carcasses still require elucidation. Therefore, we aimed to establish the causal relationship among boar taint compounds (androstenone, skatole and indole) measured in adipose tissue from pig biopsies and carcasses.

### **2.3. Materials and methods**

The data used for this study were obtained as part of routine data recording in a commercial breeding program. Samples collected for DNA extraction were only used for routine diagnostic purpose of the breeding program. Data recording and sample collection were conducted strictly in line with the rules given by Dutch Animal Research Authorities.

#### ***Data***

Data of boar taint compounds (androstenone, skatole and indole) from a Duroc-based sire line were used in the present study. These animals belonged to a breeding program whose target trait are mainly for growth and production, however, boar taint was not the target of selection. The effects of this breeding program on boar taint compounds may be considered no significant. The animals were slaughtered at approximately 177 ( $\pm 9.9$ ) days of age and boar taint compounds were measured in fat samples from the neck collected at the left carcass side, as described in Mathur et al. (2014). Boar taint compounds were also measured in fat tissue samples from biopsies, which were collected at about 302 ( $\pm 139.44$ ) days of age. Briefly, androstenone concentration was determined using liquid chromatography-mass spectrometry (Verheyden et al., 2007), whereas indole and skatole contents were measured using fluorescence at 285 and 340 nm (Ampuero Kragten et al., 2011).

Boar taint compounds levels were submitted to logarithmic transformation (log), since the variables approximately followed log-normal distributions (Duijvesteijn et al., 2010; Mathur et al., 2014). Therefore, phenotypic information consisted of the log of androstenone (AC), skatole (SC) and indole (IC) levels measured in adipose tissue of 3,590 pig carcasses and of the log of androstenone (AB), skatole (SB) and indole (IB) levels measured in adipose tissue of 397 biopsies. Most of the biopsies were performed in different animals from those carcass sampling, existing only 17 animals with biopsy and carcasses observations. The total number of animals in the pedigree was 6,401 over four generations.

#### ***Searching for phenotypic causal structures***

As described by Valente et al. (2010), SEM fitting was performed in three steps. Firstly, the posterior distributions of residual (co)variances were obtained fitting a MTM. In the second step, partial residual correlations were computed from the residual (co)variance distributions and used to build the causal structural network graphs under a high probability distribution

(HPD) interval using the ICA. Finally, the identified causal structures were incorporated in the MTM, configuring the SEM.

### *Fitting a Bayesian multi-trait model*

A Bayesian MTM was fitted according to the following mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

wherein:  $\mathbf{y}$  is a vector containing the logarithm of boar taint compounds in biopsies and carcasses;  $\mathbf{X}$  is the incidence matrix of systematic effects;  $\boldsymbol{\beta}$  is a vector of systematic effects, containing the contemporary group (farm-year-month of slaughter), the covariates age at sampling and scaled hot carcass weight at slaughter;  $\mathbf{Z}$  is the incidence matrix of animal additive genetic effects;  $\mathbf{a}$  is a vector of additive genetic effects,  $\mathbf{a} \sim N(0, \mathbf{G}_0 \otimes \mathbf{A})$ ;  $\mathbf{e}$  is a vector of residual effects,  $\mathbf{e} \sim N(0, \mathbf{R}_0 \otimes \mathbf{I})$ ;  $\mathbf{G}_0$  is the additive genetic (co)variance matrix;  $\mathbf{A}$  is the pedigree-based relationship matrix;  $\mathbf{R}_0$  is the residual (co)variance matrix;  $\mathbf{I}$  is an identity matrix.

A chain with 2,000,000 iterations was generated considering the following prior distribution for model parameters:

$$\rho(\boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{R}_0) = \rho(\boldsymbol{\beta})\rho(\mathbf{u}|\mathbf{G}_0)\rho(\mathbf{G}_0)\rho(\mathbf{R}_0) \propto N(\mathbf{u}|0, \mathbf{G}_0 \otimes \mathbf{A}) \times IW(\mathbf{G}_0|v_g, \mathbf{S}_g) \times IW(\mathbf{R}_0|v_r, \mathbf{S}_r),$$

wherein  $\rho(\boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \mathbf{R}_0)$  is the joint prior distribution assumed for the MTM;  $\rho(\boldsymbol{\beta})$  is the  $\boldsymbol{\beta}$  prior distribution;  $\rho(\mathbf{u}|\mathbf{G}_0)$  is the  $\mathbf{u}$  prior distribution conditioned to  $\mathbf{G}_0$ ;  $\rho(\mathbf{G}_0)$  is the  $\mathbf{G}_0$  prior distribution;  $\rho(\mathbf{R}_0)$  is the  $\mathbf{R}_0$  prior distribution;  $N(\mathbf{u}|0, \mathbf{G}_0 \otimes \mathbf{A})$  is a multivariate normal density centered at 0 and covariance matrix  $\mathbf{G}_0 \otimes \mathbf{A}$ ;  $IW(\mathbf{G}_0|v_g, \mathbf{S}_g)$  is an Inverse Wishart density with  $v_g$  degrees of freedom and scale matrix  $\mathbf{S}_g$ ;  $IW(\mathbf{R}_0|v_r, \mathbf{S}_r)$  is an Inverse Wishart density with  $v_r$  degrees of freedom and scale matrix  $\mathbf{S}_r$ . Uniform distribution was assigned as priori for  $\boldsymbol{\beta}$ .

A burn-in of 100,000 and thin of 1,000 iterations were used. The analysis was conducted using the BLUPF90 family of programs (Miszta et al., 2015). The convergence was verified using Geweke diagnostic criteria from the postgibbsf90 software (Miszta et al., 2015) and visual inspection of posterior distributions.

### ***Building a causal structural network***

In order to obtain causal network graphs among the traits, the posterior distributions of residual (co)variances ( $\mathbf{R}_0$ ) generated from MTM were used as input for ICA adapted for quantitative analysis by Valente et al. (2012). At the first analyses, the script written in R (R Core Team, 2017) estimate the partial correlations, which in turn were used in ICA following three steps:

**Step 1:** The partial correlations were used to identify the non-directed connection between two adjacent traits. If partial correlations between two traits were different from zero, an edge was created between them. The output of this step is a graph with traits linked (ex.:  $y_1-y_2$ ), however, the edges are not directed;

**Step 2:** The correlations were also used to identify the edge directions (ex.:  $y_1 \rightarrow y_2$ ) and collider traits. If two non-adjacent traits (ex.:  $y_1$  and  $y_3$ ) present a partial correlation and share a common trait ( $y_2$ ), the edges are oriented to the common trait (collider) (ex:  $y_1 \rightarrow y_2 \leftarrow y_3$ );

**Step 3:** When possible, non-directed edges were oriented to new unshielded colliders. This step was only necessary if the graph obtained in step 2 showed non-oriented edges.

In summary, each graph variable represent a vertex, they are connected to others by edges that indicate a causal association. This edges may be directed edges when the edges have arrows in one extremity; symmetrical direct edges, when the edges have arrows in both extremities; and undirected edges when the edges have no arrows.

As described by Valente et al. (2011), the application of the ICA involves a set of statistical decisions about declaring partial correlations as null or not using a HPD interval. Different HPD may indicate the edges and the structures that are more stable, therefore, different HPD intervals (95, 90, 85, 80, 75 and 70) were evaluated for the identification of the best network.

### ***Fitting the structural equations models***

The oriented networks returned by ICA were used to build a SEM, as presented by Gianola and Sorensen (2004). The SEM was fitted as a MTM in which causal parental variable (ex.: in the graph  $y_1 \rightarrow y_2$ ,  $y_1$  is the causal parental trait and  $y_2$  is the child trait) was considered as co-variable in each equation attributed to the child traits. In addition, the residual (co)variance matrix was diagonal ( $\Psi_0$ ). The structural equations were fitted according to the following model:

$$\mathbf{y} = (\mathbf{A} \otimes \mathbf{I})\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

in this model,  $\mathbf{y}$  vector,  $\mathbf{Z}$  and  $\mathbf{X}$  matrices are described as in the MTM;  $\boldsymbol{\beta}$  is a vector of systematic effects, containing the contemporary group (farm-year-month of slaughter), the covariates age at sampling, scaled hot carcass weight at slaughter and the parental variables;  $\mathbf{a}$  is a vector of additive genetic effects,  $\mathbf{a} \sim N(0, \mathbf{G}_0^* \otimes \mathbf{A})$ ;  $\mathbf{e}$  is a vector of residual effects,  $\mathbf{e} \sim N(0, \mathbf{R}_0^* \otimes \mathbf{I})$ ;  $\mathbf{I}$  is an identity matrix;  $\mathbf{A}$  is a square matrix with dimension equivalent to the number of traits under evaluation, the  $\mathbf{A}$  diagonal has zeros and the structural coefficients are found out of  $\mathbf{A}$  diagonal;  $\mathbf{G}_0^*$  is the SEM genetic (co)variance matrix given by  $\mathbf{G}_0^* = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{G}_0(\mathbf{I} - \mathbf{A})^{-1}$ ;  $\mathbf{R}_0^*$  is the residual diagonal variance matrix given by  $\mathbf{R}_0^* = (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\Psi}_0(\mathbf{I} - \mathbf{A})^{-1}$ ;  $\boldsymbol{\Psi}_0$  the SEM residual (co)variance matrix.

The analysis was performed considering the following prior distribution for model parameters:

$$\rho(\mathbf{A}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Psi}_0) = \rho(\mathbf{A})\rho(\boldsymbol{\beta})\rho(\mathbf{u}|\mathbf{G}_0)\rho(\mathbf{G}_0)\rho(\boldsymbol{\Psi}_0) \propto N(\mathbf{u}|0, \mathbf{G}_0 \otimes \mathbf{A}) \times IW(\mathbf{G}_0|v_G, \mathbf{G}_0^*) \times \prod_{j=1}^t Inv. \mathcal{X}^2(\psi_j|v_\psi, s^2),$$

in which  $\rho(\mathbf{A}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Psi}_0)$  is the joint prior distribution was assumed for the SEM;  $\rho(\mathbf{A})$  is the  $\mathbf{A}$  prior distribution;  $\rho(\boldsymbol{\beta})$  is the  $\boldsymbol{\beta}$  prior distribution;  $\rho(\mathbf{u}|\mathbf{G}_0)$  is the  $\mathbf{u}$  prior distribution conditioned to  $\mathbf{G}_0$ ;  $\rho(\mathbf{G}_0)$  is the  $\mathbf{G}_0$  prior distribution;  $\rho(\boldsymbol{\Psi}_0)$  is the  $\boldsymbol{\Psi}_0$  prior distribution;  $N(\mathbf{a}|0, \mathbf{G}_0 \otimes \mathbf{A})$  is a multivariate normal density centered at 0 and covariance matrix  $\mathbf{G}_0 \otimes \mathbf{A}$ ;  $IW(\mathbf{G}_0|v_G, \mathbf{G}_0^*)$  is an Inverse Wishart density with  $v_G$  degrees of freedom and scale matrix  $\mathbf{G}_0^*$ ;  $Inv. \mathcal{X}^2(\psi_j|v_\psi, s^2)$  is a scaled inverse-chi-square distribution with  $v_\psi$  degrees of freedom and scale parameter  $s^2$ ;  $\psi_j$  is the residual variance for trait  $j$ . Uniform distribution were assigned as priori for  $\boldsymbol{\beta}$  and for each structural coefficient in  $\mathbf{A}$ .

Gibbs sampler algorithm was used to obtain the posterior chain through the BLUPF90 family of programs (Misztal et al., 2015). Chains of 2,000,000 iterations, burn-in of 100,000 and thin of 1,000 iterations were used. The convergence was verified using Geweke diagnostic criteria from the postgibbsf90 software (Misztal et al., 2015) and visual inspection of the posterior distributions.

### ***Model comparison***

As different HPD intervals were used to define the causal structures the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) was applied to verify the models

goodness-of-fit. Considering  $\theta$  as a vector containing the model parameters, the DIC was obtained as follows:

$$DIC = D(\bar{\theta}) + 2p_D,$$

in which  $D(\bar{\theta})$  is the likelihood-based deviance estimate of the evaluated model and  $p_D$  is the effective number of parameters in the model. The smallest DIC value implies on the best fitting. However, DIC only expresses whether one model presented the best goodness-of-fit compared with other models, being the magnitude of this difference subjective. In order to complement this information, the Model Posterior Probabilities (MPP) presented by Wilberg and Bence (2008), were calculated as:

$$p(M_t|\theta) = \frac{\exp\left(-\frac{\Delta_t}{2}\right)}{\sum_{t=1}^4 \exp\left(-\frac{\Delta_t}{2}\right)},$$

in which  $t$  are the different models obtained considering the causal structure returned;  $p(M_t|\theta)$  is the posterior probability of model  $t$  be the best among the set of models compared and  $\Delta_t$  is the DIC difference between model  $t$  and the model that presented the smallest DIC value. The  $\Delta_t$  for the model with the smallest DIC value is equal to zero.

## 2.4. Results

### *Multi-trait variance components*

The posterior means and 95% HPD intervals for residual and genetics variances for each trait, as well as the genetic and residual correlations between traits from the first step of the analysis (fitting an MTM), are shown in Table 1.

High positive residual correlations were observed between AC-SB, AC-IB and SC-IC. Null residual correlations were observed between IC-AB, IC-IB and AB-SB. High positive genetic correlations were observed between SC-IC, SC-AB, SC-SB and IC-AB. Moreover, null genetic correlations were observed between AC-SB, AC-IB, SC-IB, IC-SB, IC-IB, AB-SB, AB-IB and SB-IB. We also found moderate genetic and residual correlations between AC-AB, SC-SB, IC-IB.

Table 1 Posterior means and 95% high probability distribution (HPD) intervals for the dispersion parameters fitting a multi-trait model

| Parameter <sup>1</sup> | Posterior mean | 95% Interval   | HPD | Parameter <sup>2</sup> | Posterior mean | 95% Interval  | HPD |
|------------------------|----------------|----------------|-----|------------------------|----------------|---------------|-----|
| $\sigma_{eAC}^2$       | 0.87           | [0.80, 0.95]   |     | $\sigma_{gAC}^2$       | 0.27           | [0.19, 0.37]  |     |
| $r_{eACeSC}$           | 0.39           | [0.33, 0.45]   |     | $r_{gACgSC}$           | 0.28           | [0.06, 0.48]  |     |
| $r_{eACeIC}$           | 0.38           | [0.33, 0.43]   |     | $r_{gACgIC}$           | 0.26           | [0.03, 0.46]  |     |
| $r_{eACeAB}$           | 0.53           | [0.27, 0.74]   |     | $r_{gACgAB}$           | 0.48           | [0.04, 0.81]  |     |
| $r_{eACeSB}$           | 0.82           | [0.62, 0.94]   |     | $r_{gACgSB}$           | 0.13           | [-0.46, 0.65] |     |
| $r_{eACeIB}$           | 0.61           | [0.25, 0.86]   |     | $r_{gACgIB}$           | -0.40          | [-0.79, 0.17] |     |
| $\sigma_{eSC}^2$       | 0.25           | [0.22, 0.28]   |     | $\sigma_{gSC}^2$       | 0.14           | [0.10, 0.18]  |     |
| $r_{eSceIC}$           | 0.68           | [0.64, 0.72]   |     | $r_{gSCgIC}$           | 0.69           | [0.56, 0.81]  |     |
| $r_{eSceAB}$           | -0.34          | [-0.63, -0.02] |     | $r_{gSCgAB}$           | 0.69           | [0.31, 0.92]  |     |
| $r_{eSceSB}$           | 0.41           | [0.09, 0.69]   |     | $r_{gSCgSB}$           | 0.59           | [0.30, 0.89]  |     |
| $r_{eSceIB}$           | 0.48           | [0.08, 0.78]   |     | $r_{gSCgIB}$           | 0.42           | [-0.01, 0.78] |     |
| $\sigma_{eIC}^2$       | 0.19           | [0.18, 0.21]   |     | $\sigma_{gIC}^2$       | 0.07           | [0.05, 0.09]  |     |
| $r_{eICeAB}$           | -0.17          | [-0.47, 0.16]  |     | $r_{gICgAB}$           | 0.73           | [0.37, 0.93]  |     |
| $r_{eICeSB}$           | 0.52           | [0.18, 0.78]   |     | $r_{gICgSB}$           | 0.09           | [-0.43, 0.55] |     |
| $r_{eICeIB}$           | 0.31           | [-0.15, 0.70]  |     | $r_{gICgIB}$           | 0.34           | [-0.11, 0.74] |     |
| $\sigma_{eAB}^2$       | 0.26           | [0.15, 0.39]   |     | $\sigma_{gAB}^2$       | 0.44           | [0.29, 0.62]  |     |
| $r_{eABeSB}$           | 0.19           | [-0.07, 0.46]  |     | $r_{gABgSB}$           | 0.36           | [-0.01, 0.71] |     |
| $r_{eABeIB}$           | 0.34           | [0.06, 0.61]   |     | $r_{gABgIB}$           | 0.29           | [-0.04, 0.74] |     |
| $\sigma_{eSB}^2$       | 0.76           | [0.56, 1.00]   |     | $\sigma_{gSB}^2$       | 0.38           | [0.16, 0.67]  |     |
| $r_{eSBeIB}$           | 0.47           | [0.24, 0.67]   |     | $r_{gSBgIB}$           | 0.37           | [-0.04, 0.74] |     |
| $\sigma_{eIB}^2$       | 0.58           | [0.38, 0.81]   |     | $\sigma_{gIB}^2$       | 0.53           | [0.28, 0.81]  |     |

<sup>1</sup>  $\sigma_{e_i}^2$  = residual variance of trait  $i$ ,  $i=AC$  (androstenone in carcass),  $SC$  (skatole in carcass),  $IC$  (indole in carcass)  $AB$  (androstenone in biopsies),  $SB$  (skatole in biopsies),  $IB$  (indole in biopsies),  $r_{e_i e_{i'}}$  = residual correlation between traits  $i$  and  $i'$ ;

<sup>2</sup>  $\sigma_{g_i}^2$  = additive genetic variance of trait  $i$ ,  $r_{g_i g_{i'}}$  = additive genetic correlation between traits  $i$  and  $i'$ .

The posterior distributions of the heritabilities obtained from the MTM are presented in Figure 1.

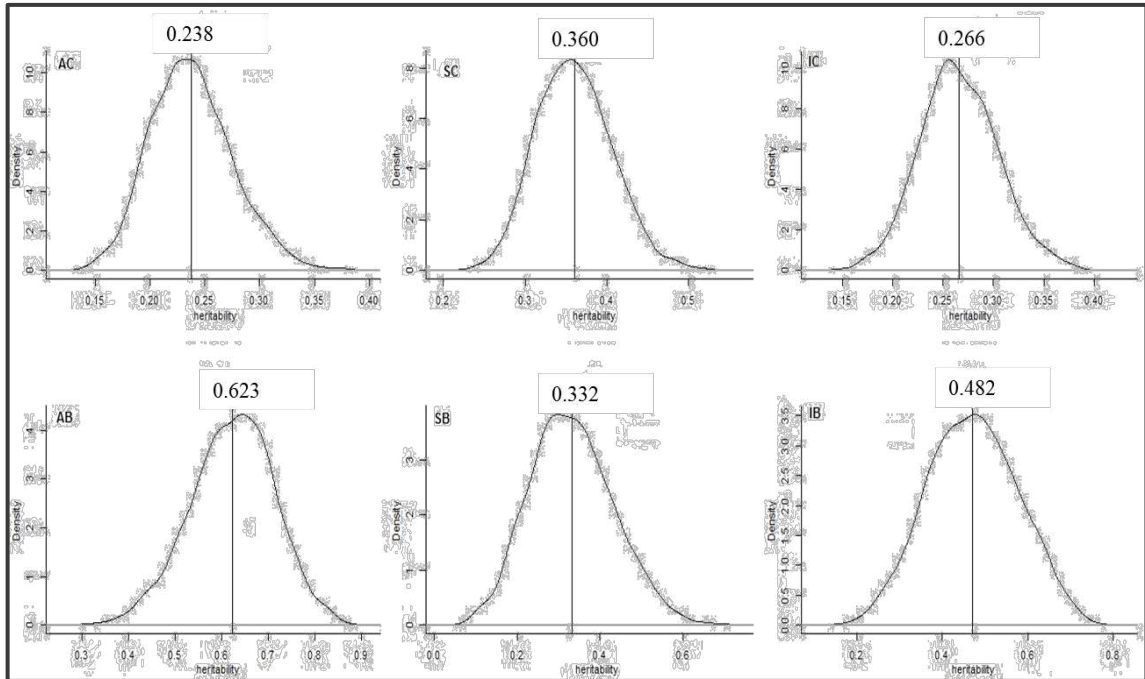


Figure 1 Posterior distributions of heritabilities from the multi-trait model. AC = androstenone in carcass; SC = skatole in carcass; IC = indole in carcass; AB = androstenone in biopsies; SB = skatole in biopsies, IB = indole in biopsies.

The heritabilities of boar taint compounds measured in carcasses ranged from 0.238 to 0.360, whereas for boar taint compounds measured in biopsies the heritabilities were higher, ranging from 0.332 to 0.623.

### *Causal structures identification*

Four different graphs were returned, considering 95%, 90% and 85% (Figure 2a-c, respectively), from 80 to 70% HPD intervals the same structure was obtained (Figure 2d).

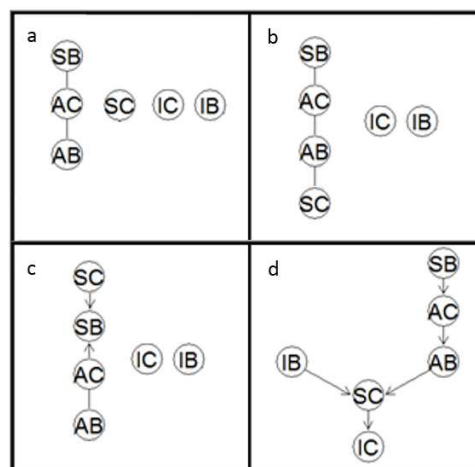


Figure 2 Graphs returned by the Inductive Causation algorithm considering high probability distribution intervals of 95% (a), 90% (b), 85% (c) and 80 to 70% (d) for the statistical decisions



involving the studied traits: AC = androstenone in carcass; SC = skatole in carcass; IC = indole in carcass; AB = androstenone in biopsies; SB = skatole in biopsies, IB = indole in biopsies.

Undirected graphs were returned considering 95% and 90% HPD intervals and their structures were used as ‘skeleton’ to create oriented graphs applying priori information. We oriented their edges considering that boar taint compounds measured in biopsies may be used to predict the compounds measured in carcasses, since biopsies are realized in live animals, before slaughter, therefore consisting of previous observations exerting effects on subsequent (carcass) observations, as presented in Figure 3.

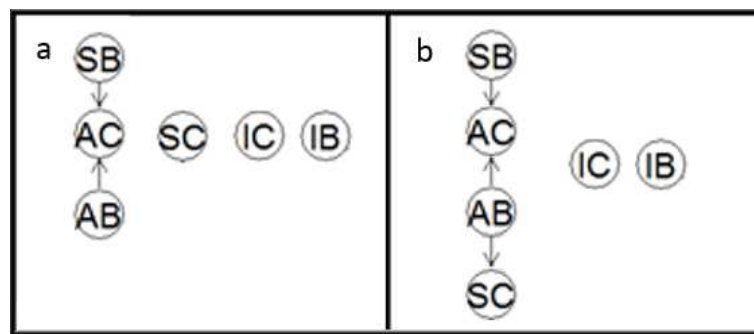


Figure 3 Graphs built combining prior information and Inductive Causation algorithm output considering high probability distribution intervals of 95% (a) and 90% (b) for the statistical decisions involving the studied traits: AC = androstenone in carcass; SC = skatole in carcass; IC = indole in carcass; AB = androstenone in biopsies; SB = skatole in biopsies, IB = indole in biopsies.

The observed causal structures considering 85% (Figure 2c) and 80 to 70% HPD intervals (Figure 2d) were used to build SEM<sub>85</sub> and SEM<sub>80</sub>, respectively, and the causal structures presented in Figures 3a and 3b were used to build SEM<sub>95P</sub> and SEM<sub>90P</sub>, respectively.

### ***Comparing models***

Four distinct SEM were constructed conditionally on the causal structures presented in Figures 2c, 2d, 3a and 3b. All SEMs carried unshielded colliders that were supported by data evidence, i.e. their presence in the causal structure may improved the models fitting. The goodness of fit was verified using DIC values obtained for each studied model (Table 2) as well as the MPP comparing the models to SEM<sub>90P</sub> (MPP<sub>1</sub>) and to SEM<sub>80</sub> (MPP<sub>2</sub>)

Table 2 Models comparison

| Model                          | DIC <sup>1</sup> | MPP <sub>1</sub> | MPP <sub>2</sub> |
|--------------------------------|------------------|------------------|------------------|
| MTM <sup>2</sup>               | 61,061.67        | ≈0.00            | ≈0.00            |
| SEM <sub>85</sub> <sup>3</sup> | 1,831.66         | ≈0.00            | ≈0.00            |
| SEM <sub>80</sub>              | 1,682,45         | ≈0.00            | ≈0.99            |
| SEM <sub>95P</sub>             | 2,274.78         | ≈0.00            | ≈0.00            |
| SEM <sub>90P</sub>             | 1,632.92         | ≈1.00            | -                |

<sup>1</sup> Deviance Information Criterion values obtained for each model;

<sup>2</sup> Multiple Trait Model (MTM)

<sup>3</sup> Structural Equations Models (SEM<sub>85</sub>, SEM<sub>80</sub>, SEM<sub>95P</sub>, and SEM<sub>90P</sub>)

<sup>4</sup> Model Posterior Probabilities comparing the models to SEM<sub>90P</sub>;

<sup>5</sup> Model Posterior Probabilities comparing the models to SEM<sub>80</sub>.

The model with best goodness-of-fit was SEM<sub>90P</sub>, elaborated considering a priori information; the second model with best model was SEM<sub>80</sub>, elaborated considering only IC output (Table 2). The SEM<sub>90P</sub> was the most probable model that presented the best goodness-of-fit (MPP<sub>1</sub> ≈1.00; Table 2), however, in posterior analyses, this model showed several unexpected null correlations (Table 3) therefore, the results returned by the second model were also evaluated. In this sense, we calculate the MPP<sub>2</sub>, in which SEM<sub>90P</sub> was considered and the model with best goodness-of-fit was the SEM<sub>80</sub>. In this analysis, the SEM<sub>80</sub> was the most probable model that presented the best goodness-of-fit (MPP<sub>1</sub> ≈0.99; Table 2).

The posterior means and 95% HPD intervals for each trait genetic and residual variances and correlations obtained fitting SEM<sub>80</sub> and SEM<sub>90P</sub> are shown in Table 3.

Table 3 Posterior means and 95% high probability distribution (HPD) intervals for the dispersion parameters fitting SEM<sub>80</sub> and SEM<sub>90P</sub>.

| SEM <sub>80</sub> <sup>1</sup> |                |                  | SEM <sub>90P</sub> <sup>2</sup> |                |                  |
|--------------------------------|----------------|------------------|---------------------------------|----------------|------------------|
| Parameter <sup>3</sup>         | Posterior mean | 95% HPD Interval | Parameter <sup>3</sup>          | Posterior mean | 95% HPD Interval |
| $\psi_{AC}$                    | 0.70           | [0.64, 0.76]     | $\psi_{AC}$                     | 0.68           | [0.62, 0.74]     |
| $\psi_{SC}$                    | 0.08           | [0.06, 0.10]     | $\psi_{SC}$                     | 0.07           | [0.05, 0.09]     |
| $\psi_{IC}$                    | 0.07           | [0.06, 0.09]     | $\psi_{IC}$                     | 0.07           | [0.06, 0.09]     |
| $\psi_{AB}$                    | 0.26           | [0.13, 0.38]     | $\psi_{AB}$                     | 0.28           | [0.17, 0.41]     |
| $\psi_{SB}$                    | 0.31           | [0.18, 0.48]     | $\psi_{SB}$                     | 0.44           | [0.23, 0.63]0    |
| $\psi_{IB}$                    | 0.39           | [0.23, 0.58]     | $\psi_{IB}$                     | 0.45           | [0.29, 0.58]     |
| $\sigma_{gAC}^2$               | 0.56           | [0.47, 0.66]     | $\sigma_{gAC}^2$                | 0.59           | [0.50, 0.70]     |
| $r_{gACgSC}$                   | 0.72           | [0.66, 0.79]     | $r_{gACgSC}$                    | 0.71           | [0.65, 0.77]     |
| $r_{gACgIC}$                   | 0.75           | [0.69, 0.80]     | $r_{gACgIC}$                    | 0.73           | [0.67, 0.79]     |
| $r_{gACgAB}$                   | 0.61           | [0.34, 0.80]     | $r_{gACgAB}$                    | 0.09           | [-0.43, 0.60]    |
| $r_{gACgSB}$                   | 0.76           | [0.63, 0.87]     | $r_{gACgSB}$                    | -0.10          | [-0.73, 0.52]    |
| $r_{gACgIB}$                   | 0.22           | [-0.08, 0.54]    | $r_{gACgIB}$                    | -0.26          | [-0.62, 0.12]    |
| $\sigma_{gSC}^2$               | 0.45           | [0.41, 0.49]     | $\sigma_{gSC}^2$                | 0.46           | [0.42, 0.51]     |
| $r_{gSCgIC}$                   | 0.92           | [0.89, 0.94]     | $r_{gSCgIC}$                    | 0.92           | [0.89, 0.94]     |
| $r_{gSCgAB}$                   | 0.41           | [0.18, 0.65]     | $r_{gSCgAB}$                    | -0.01          | [-0.53, 0.49]    |
| $r_{gSCgSB}$                   | 0.87           | [0.74, 0.96]     | $r_{gSCgSB}$                    | 0.22           | [-0.43, 0.85]    |
| $r_{gSCgIB}$                   | 0.53           | [0.22, 0.79]     | $r_{gSCgIB}$                    | 0.24           | [-0.13, 0.61 ]   |
| $\sigma_{gIC}^2$               | 0.28           | [0.25, 0.31]     | $\sigma_{gIC}^2$                | 0.28           | [0.26, 0.31]     |
| $r_{eICeAB}$                   | 0.39           | [0.08, 0.69]     | $r_{eICeAB}$                    | 0.01           | [-0.56, 0.58]    |
| $r_{gICgSB}$                   | 0.70           | [0.53, 0.84]     | $r_{gICgSB}$                    | -0.01          | [-0.67, 0.71]    |
| $r_{gICgIC}$                   | 0.44           | [0.09, 0.75]     | $r_{gICgIC}$                    | 0.06           | [-0.33, 0.50]    |
| $\sigma_{gAB}^2$               | 0.47           | [0.29, 0.66]     | $\sigma_{gAB}^2$                | 0.42           | [0.23, 0.59]     |
| $r_{gABgSB}$                   | 0.44           | [0.28, 0.62]     | $r_{gABgSB}$                    | 0.43           | [0.23, 0.62]     |
| $r_{gABgIB}$                   | 0.51           | [0.33, 0.70]     | $r_{gABgIB}$                    | 0.57           | [0.43, 0.71]     |
| $\sigma_{gSB}^2$               | 1.14           | [0.91, 1.38]     | $\sigma_{gSB}^2$                | 0.82           | [0.49, 1.14]     |
| $r_{gSBgIB}$                   | 0.61           | [0.47, 0.75]     | $r_{gSBgIB}$                    | 0.72           | [0.55, 0.86]     |
| $\sigma_{gIB}^2$               | 0.79           | [0.53, 1.08]     | $\sigma_{gIB}^2$                | 0.73           | [0.52, 0.95]     |

<sup>1</sup> Structural Equations Models (SEM) considering causal structures obtained with 80-70% high probability distribution (HPD) interval;

<sup>2</sup> SEM) considering causal structures obtained with 90% high probability distribution HPD interval plus priori;

<sup>3</sup>  $\psi_i$ = residual variance of trait i.  $i$  =AC (androstenone in carcass), SC (skatole in carcass), IC (indole in carcass), AB (androstenone in biopsies), SB (skatole in biopsies), IB (indole in

biopsies),  $\sigma_{g_i}^2$  = additive genetic variance of trait  $i$ ,  $r_{g_i g_{i'}}$  = additive genetic correlation between traits  $i$  and  $i'$ .

The residual ( $\psi_i$ ) and genetic ( $\sigma_{g_i}^2$ ) variance estimates obtained by both SEM for all traits may be considered equal since there is a large overlap of their HPD intervals. In addition, regarding the HPD contents, the SEM<sub>90P</sub> returned null genetic correlations between most of boar taint compounds. On the other hand, in SEM<sub>80</sub>, only the genetic correlation between AC-IB was considered null. Moreover, the SEM<sub>90P</sub> provided genetic correlations between the same boar taint compounds measured in carcasses and biopsies null and with large HPD interval, while in SEM<sub>80</sub> these genetic correlations were positive and with small HPD interval.

The posteriori means of structural coefficients pertaining to both SEM<sub>80</sub> and SEM<sub>90P</sub> are presented in Table 4.

Table 4 Structural coefficients ( $\lambda$ ) pertaining to SEM<sub>80</sub> and SEM<sub>90P</sub> models.

| Structural coefficient         | SEM <sub>90P</sub> <sup>1</sup> |                 | SEM <sub>80</sub> <sup>2</sup> |         |
|--------------------------------|---------------------------------|-----------------|--------------------------------|---------|
|                                | Mean <sup>2</sup>               | SD <sup>3</sup> | mean                           | SD      |
| $\lambda_{SC,AB}$ <sup>4</sup> | -0.0002                         | 0.0001          | -0.0006                        | 0.0002  |
| $\lambda_{AC,AB}$              | 0.0007                          | 0.0004          | -                              | -       |
| $\lambda_{AC,SB}$              | -0.0006                         | 0.0003          | 0.0002                         | 0.0002  |
| $\lambda_{AB,AC}$              | -                               | -               | -0.0002                        | 0.0002  |
| $\lambda_{SC,IB}$              | -                               | -               | 0.0004                         | 0.0002  |
| $\lambda_{IC,SC}$              | -                               | -               | 0.0088                         | 0.00129 |

<sup>1</sup> Structural Equations Models (SEM) considering causal structures obtained with 80-70% high probability distribution (HPD) interval;

<sup>2</sup> SEM considering causal structures obtained with 90% high probability distribution HPD interval plus priori;

<sup>2</sup> estimated mean of structural coefficient;

<sup>3</sup> standard deviation;

<sup>4</sup> AC: androstenone in carcass, SC: skatole in carcass, IC: indole in carcass, AB: androstenone in biopsies, SB: skatole in biopsies, IB: indole in biopsies.

Regarding the structural coefficients obtained with SEM<sub>90P</sub>, we can infer that AB has a negative effect on SC and a positive effect on AC, while SB has a negative effect on AC. On the other hand, the structural coefficients obtained with SEM<sub>80</sub> indicate that SB has a positive effect on AC, IB has a positive effect on SC and SC has a positive effect on IC, whereas AB has a negative effect on SC and AC has a negative effect on AB.

## 2.5. Discussion

In this research a multi-trait approach were used to evaluate the boar taint compounds considering carcasses and biopsies measurements as different traits. In general, boar taint compounds presented greater phenotypic variation in biopsies than in carcasses; nevertheless, most part of this variation was represented by the genetic variance. For the traits measured in carcasses, the genetic to phenotypic variance rate was slightly smaller, resulting in moderate heritabilities, while for the traits measured in biopsies, higher heritabilities were found. Our findings are in agreement with previously reported boar taint compounds heritabilities, which ranged from 0.46 to 0.72, 0.50 to 0.26 and 0.29 to 0.35 for androstenone, skatole and indole measured in carcasses, respectively (Campos et al., 2015; Mathur et al., 2012; Windig et al., 2012). The genetic variance observed for boar taint traits measured in the biopsies were greater than the observed for carcass measured traits. Most of biopsies information belongs to older animals (302 days-old, on average) than to those used to measure the carcass boar taint compounds (177 days-old, on average). The differences among heritabilities and genetic variance observed in carcasses and biopsies indicate that higher genetic gain can be obtained using biopsy measurements. However, this finding should be confirmed using larger dataset, since large HPD intervals were observed for biopsy parameters.

The genetic correlations between boar taint compounds measured in carcasses are well known and described as favorable and positive (Campos et al., 2015; Mathur et al., 2012; Tajet et al., 2006; Windig et al., 2012; Zamaratskaia et al., 2004). However, we have found no studies reporting the genetic correlations between carcasses and biopsies measurements. The genetic correlations among boar taint compounds in biopsies and carcasses found the present study suggest that biopsies and carcass measurements may be considered different traits. This may justified by difference in the average age at measurements on carcasses and biopsies. The biopsies were performed in older animals, which already are in reproductive phase, producing sexual hormones in the testis, including androstenone that affect the other boar taint compounds. At the average age of carcass measurements, the animals may be at the beginning of reproductive phase and not present a full testis activity, presenting phenotypes different from those as adults.

The MTM analyses provided residual correlations between boar taint compounds, which were used as input for the ICA to search for causal structures based on different HPD intervals. Since we have used several non-null residual correlations as input for ICA, it was able to return partially directed or fully directed causal graphs. The edges orientation were supported by the

existence of unshielded colliders, which were essential for the orientation of these edges by the algorithm (Pearl, 2003; Valente et al., 2011, 2010). Nevertheless, the directed graphs were obtained using 85% or less HPD interval. Undirected graphs returned with 95 and 90% HPD intervals were be oriented combining the ICA output and the prior knowledge (Pearl, 2003; Valente et al., 2011).

Small HPD intervals make the analysis more parsimonious (less restrictive / strict), enabling the identification of different causal structures and new colliders (Valente et al., 2010) that may impose and change the edges direction. We observed changes in edge direction between SB and AC (Figure 1c and 1d), which may be justified by the identification of new unshielded colliders that redirected the graph and formed a pathway (Pearl, 2003). It seems logical that androstenone is the parental variable for skatole due to androstenone antagonism in skatole degradation in liver (Zamaratskaia and Squires, 2009). However, biochemically, possible feedback between androstenone and skatole has also been reported (Claus et al., 1994; Rius et al., 2005; Zamaratskaia et al., 2004). Therefore, the causal relationship indicating the effects of skatole on androstenone is also plausible, since edges may have alternative directions without contradict known biochemical paths (Bouwman et al., 2014). In addition, the approach used allows the identification of causal relationships despite the genetic effects using the phenotypes conditional distribution (Valente et al., 2011, 2010).

The plausibility of all network structures obtained with different HPD intervals was verified before fitting the SEM, which incorporated the parental variables into the prediction equations of the child variables, resulting in four different SEM that were compared by the DIC. The model with the best goodness-of-fit was SEM<sub>90P</sub>, followed by SEM<sub>80</sub>, and the simplest model (MTM) presented the worst goodness-of-fit among all models. The DIC does not assign better scores to complex models if the extra goodness-of-fit achieved do not compensate the increase in the number of parameters (Valente et al., 2011). Moreover, MPP showed that SEM<sub>90P</sub> presented the best goodness-of-fit compared to the other models. Therefore, even being more parameterized, all SEMs presented better goodness-of-fit than MTM. Based on the given causal structure, this indicates that the variability of each boar taint compound can be partially explained by the conditioning (parent) boar taint compound resulting in a model that is more parsimonious than MTM.

The differences observed in genetic and residual variance among MTM and SEM may be attributed to the model re-parameterization that decreased residual variances and increased genetic variances. In this sense, the interpretation of variance components obtained by MTM

and SEM should not be the same (Valente et al., 2013), since models specific genetic (co)variances refer to the (co)dispersion of the genetic effects of each model, and therefore have distinct meanings (Bouwman et al., 2014).

Conditioning of correlated traits in SEM resulted in greater genetic correlations than those obtained in MTM. These posterior means of genetic correlations from SEM refer to the genetic covariance that remains after conditioning on the appropriate boar taint compound. In other words, genetic correlations estimated from SEM express the causal correlation between direct genetic effects for each trait (Bouwman et al., 2014); therefore, all sources of association among traits are accounted for by the SEM via the causal connections among phenotypes (Valente et al., 2013).

The posterior means of residual and genetic variances estimated in SEM<sub>80</sub> and SEM<sub>90P</sub> may be considered equal due to the overlap of their 95% HPD intervals, indicating class of equivalent causal structures (Valente et al., 2010). This was expected, since the different graphs used to build the SEM were returned by the ICA and represented classes of equivalent causal structures (Pearl, 2003). Despite the HPD overlap, the estimates were consistently more accurate in the SEM<sub>80</sub> (lower HPD interval content), suggesting better estimates.

Although high posterior means of residual and genetic variances have been estimated, several genetic correlations obtained in the SEM<sub>90P</sub> may be considered null, since the value zero is present in their 95% HPD intervals. It was expected that the same boar taint compound (androstenone, skatole or indole) measured in biopsies and in carcasses will had positive genetic correlations, however these correlations were not different from zero with the SEM<sub>90P</sub>. These findings indicate that, despite the best goodness-of-fit of SEM<sub>90P</sub> evidenced by DIC, this model may be poorly compatible with data evidences, in terms that posterior dispersion parameters intervals were substantially large.

As previously discussed, the use of SEM allows the identification of parental traits effects on their child traits through structural coefficients (Valente et al., 2011, 2010). Nevertheless, the studied phenotypes were logarithmically transformed, therefore, it is not possible to declare the real magnitude of the effect that each parental boar taint compound exerts on its child, but the direction of these effects can be perfectly assessed.

Although unexpected, in the SEM, some negative effects were observed between some boar taint compounds ( $AB \rightarrow SC$  and  $SB \rightarrow AC$  in SEM<sub>90P</sub>;  $AB \rightarrow SC$  and  $AC \rightarrow AB$  in SEM<sub>80</sub>). This possibly occurs because the SEM enables indirect identification of a second source of genetic association in which the set of genes that affect the phenotype of the parental variable

may have negative effect on the phenotype of the child variable (Bouwman et al., 2014; Valente et al., 2013). Thus, this indirect source of covariation could present an opposite sign to that of the covariance between direct genetic effects, indicating that genes affecting a pair of traits could actually have “double consequences” (Valente et al., 2013). This is plausible since biopsies measurements are performed under different metabolism conditions from those observed in carcasses measurement (after slaughter) resulting in different genic mechanism involved in the phenotype expression.

On the other hand, either SEM<sub>90P</sub> or SEM<sub>80</sub> returned positive structural coefficients in which biopsies boar taint compounds affect carcasses boar taint compounds (AB → AC in SEM<sub>90P</sub>; SB → AC and IB → SC in SEM<sub>80</sub>). As previously described, AC and AB were considered distinct traits, however, they are the same component measured using different methodologies, therefore, a positive effect was expected. In this sense, it was also expected to observe direct effects between SC: SB and IC: IB, however the ICA did not report evidence of these causalities, probably due to the small size of the biopsy database. The effect of SB on AC may be related to possible feedback that skatole exerts on androstenone (Claus et al., 1994; Rius et al., 2005; Zamaratskaia et al., 2004). The positive effects of IB on SC as well as SC on IC, both verified in SEM<sub>80</sub>, may be perfectly justified by the similar metabolic origin, synthesis and degradation pathways of these two compounds (Aluwé et al., 2011; Claus et al., 1994; Grindflek et al., 2011; Jesen et al., 1995). Furthermore, since both compounds are chemically very similar, there may be some confusion in their measurements, especially the possibility that indole measurements are "inflated" by the skatole measurements (Verheyden et al., 2007), which would also justify the positive effects.

## 2.6. Conclusion

The multi-trait model reveals that the same boar taint compound in carcass and biopsies should not be considered the same trait. Causal structure among boar taint compounds in carcasses and biopsies were identified indicating that skatole in biopsies is the major trait that exerts effects in the other boar taint compounds. The best causal structure returned was IB → SC ← AB ← AC ← SB: SC → IC, which was obtained with 80% HPD interval, this pathway indicates that skatole in carcass exerts direct and indirect effects in the other boar taint compounds. In addition, androstenone, skatole and indole measured in biopsies exert direct effects at least on one boar taint compound measured in carcasses and are passive to be used in



breeding programs. The goodness-of-fit of the structural equation models were superior to the multi-trait model.

## 2.7. References

- Aluwé, M., S. Millet, K.M. Bekaert, F.A.M.M. Tuyttens, L. Vanhaecke, et al. 2011. Influence of breed and slaughter weight on boar taint prevalence in entire male pigs. *Animal* 5(8): 1283–1289. doi: 10.1017/S1751731111000164.
- Ampuero Kragten, S., B. Verkuylen, H. Dahlmans, M. Hortos, J.A. Garcia-Regueiro, et al. 2011. Inter-laboratory comparison of methods to measure androstenone in pork fat. *Animal* 5(10): 1634–1642. doi: 10.1017/S1751731111000553.
- Bouwman, A.C., B.D. Valente, L.L.G. Janss, H. Bovenhuis, and G.J.M. Rosa. 2014. Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. *Genet Sel Evol* 46(1): 2. doi: 10.1186/1297-9686-46-2.
- Campos, C.F. de, M.S. Lopes, F.F. e Silva, R. Veroneze, E.F. Knol, et al. 2015. Genomic selection for boar taint compounds and carcass traits in a commercial pig population. *Livest Sci* 174: 10–17. doi: 10.1016/j.livsci.2015.01.018.
- Claus, R., U. Weiler, and A. Herzog. 1994. Physiological Aspects of Androstenone and Skatole Formation in the Boar A Review with Experimental Data. *Meat Sci* 38: 289–305. doi: 10.1016/0309-1740(94)90118-X.
- Gianola, D., and D. Sorensen. 2004. Quantitative Genetic Models for Describing Simultaneous and Recursive Relationships Between Phenotypes. *Genetics* 167(3): 1407–1424. doi: 10.1534/genetics.103.025734.
- Grindflek, E., T.H.E. Meuwissen, T. Aasmundstad, H. Hamland, M.H.S. Hansen, et al. 2011. Revealing genetic relationships between compounds affecting boar taint and reproduction in pigs. *J Anim Sci* 89(3): 680–692. doi: 10.2527/jas.2010-3290.
- Jesen, M.T., R.P. Cox, and B.B. Jesen. 1995. 3-Methylindole ( Skatole ) and Indole Production by Mixed Populations of Pig Fecal Bacteria. *Appl Environ Microbiol* 61(8): 3180–3184.
- Lee, G.J., A.L. Archibald, A.S. Law, S. Lloyd, J. Wood, et al. 2005. Detection of quantitative trait loci for androstenone, skatole and boar taint in a cross between Large White and Meishan pigs. *Anim Genet* 36(1): 14–22. doi: 10.1111/j.1365-2052.2004.01214.x.
- Mathur, P.K., J. ten Napel, R.E. Crump, H.A. Mulder, and E.F. Knol. 2014. Genetic relationship between boar taint compounds, human nose scores, and reproduction traits in pigs. *J Anim Breed Genet* 91(9): 4080–4089. doi: <https://doi.org/10.2527/jas.2013-6478>.
- Mathur, P.K., J. ten Napel, S. Bloemhof, L. Heres, E.F. Knol, et al. 2012. A human nose scoring system for boar taint and its relationship with androstenone and skatole. *Meat Sci* 91(4): 414–422. doi: 10.1016/j.meatsci.2012.02.025.
- Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, et al. 2015. Manual for BLUPF90 family of programs. Univ Georg Athens, USA: 125. [http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90\\_all2.pdf](http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf).
- Pearl, J. 2003. Causality: Models, Reasoning, and Inference. *Econometric Theory*. p. 675–685

- R Core Team. 2017. R: A language and environment for statistical computing.
- Rius, M.A., M. Hortós, and J.A. García-Regueiro. 2005. Influence of volatile compounds on the development of off-flavours in pig back fat samples classified with boar taint by a test panel. *Meat Sci* 71(4): 595–602. doi: 10.1016/j.meatsci.2005.03.014.
- Schumacker, R.E., and R.G. Lomax. 2010. *A Beginner's Guide to Structural Equation Modeling* (R.G. Lomax, editor). 3rd ed. Taylor & Francis Group, New York.
- Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B (Statistical Methodol)* 64(4): 583–639. doi: 10.1111/1467-9868.00353.
- Tajet, H., Ø. Andresen, and T.H.E. Meuwissen. 2006. Estimation of genetic parameters of boar taint; skatole and androstenone and their correlations with sexual maturation. *Acta Vet Scand* 48(SUPPL.1): 2–5. doi: 10.1186/1751-0147-48-S1-S9.
- Valente, B.D., G.J.M. Rosa, D. Gianola, X.L. Wu, and K. Weigel. 2013. Is structural equation modeling advantageous for the genetic improvement of multiple traits? *Genetics* 194(3): 561–572. doi: 10.1534/genetics.113.151209.
- Valente, B.D., G.J.M. Rosa, G. de Los Campos, D. Gianola, and M.A. Silva. 2010. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics* 185(2): 633–644. doi: 10.1534/genetics.109.112979.
- Valente, B.D., G.J. Rosa, M.A. Silva, R.B. Teixeira, and R.A. Torres. 2011. Searching for phenotypic causal networks involving complex traits: an application to European quail. *Genet Sel Evol* 43(1): 37. doi: 10.1186/1297-9686-43-37.
- Verheyden, K., H. Noppe, M. Aluwé, S. Millet, J. Vanden Bussche, et al. 2007. Development and validation of a method for simultaneous analysis of the boar taint compounds indole, skatole and androstenone in pig fat using liquid chromatography–multiple mass spectrometry. *J Chromatogr A* 1174(1–2): 132–137. doi: 10.1016/j.chroma.2007.08.075.
- Wilberg, M.J., and J.R. Bence. 2008. Performance of deviance information criterion model selection in statistical catch-at-age analysis. *Fish Res* 93(1–2): 212–221. doi: 10.1016/j.fishres.2008.04.010.
- Windig, J.J., H.A. Mulder, J. ten Napel, E.F. Knol, P.K. Mathur, et al. 2012. Genetic parameters for androstenone, skatole, indole, and human nose scores as measures of boar taint and their relationship with finishing traits. *J Anim Sci* 90(7): 2120–2129. doi: 10.2527/jas.2011-4700.
- Zamaratskaia, G., J. Babol, H. Andersson, and K. Lundström. 2004. Plasma skatole and androstenone levels in entire male pigs and relationship between boar taint compounds, sex steroids and thyroxine at various ages. *Livest Prod Sci* 87(2–3): 91–98. doi: 10.1016/j.livprodsci.2003.09.022.
- Zamaratskaia, G., and E.J. Squires. 2009. Biochemical, nutritional and genetic effects on boar taint in entire male pigs. *Animal* 3(11): 1508–1521. doi: 10.1017/S1751731108003674.

## CHAPTER 3

### Applying an association weight matrix in genomic prediction of boar taint compounds

#### 3.1. Abstract

Genome-wide association studies (GWAS) have identified markers associated with boar taint compounds (androstenone, skatole, and indole), which may supply valuable information to attribute different weights for SNP markers in genome-wide selection (GWS) to increase the predictive ability. A feasible way to approach non-false positive markers could be to explore biological information and genetic correlation between traits under a gene network framework based on an association weight matrix (AWM). Therefore, we aimed to evaluate the predictive ability and the bias achieved using models that allow weighting each SNP in the G matrix according to its explained genetic variance, with and without incorporating biological information in the weighting procedure. In addition, we also aimed to evaluate the effect of different population structures (multi-line and single-line) in the GWS of boar taint compounds. Boar taint compounds measured in 4,922 pig carcasses and genotypes (Illumina PorcineSNP60 BeadChip) from 3,749 animals were available for this study. Firstly, we performed a GWAS in which a single- (SL) and a multi-line (ML) population were used to identify SNPs associated with boar taint compounds using the single-step GBLUP (ssGBLUP) method. In a second step, 1%, 2%, 5% and 10% of the markers explaining the highest proportions of the genetic variance for each trait were selected to build gene networks via the AWM approach. The total number of gene interactions for each gene in the network was used to compute weights for previously selected SNP and used to build genomic relationship matrices for ssGBLUP, which we called AWM-WssGBLUP approach. In addition, we performed weighted ssGBLUP (WssGBLUP) and ssGBLUP analyses as standard scenarios and their predictive ability were compared with the ones of the AWM-WssGBLUP approach. The WssGBLUP showed greater predictive ability for androstenone in both SL and ML scenarios compared to the other methods. For skatole and indole, the AWM-WssGBLUP method using the top 5% SNPs slightly increased the predictive ability by up to 4% compared to the traditional ssGBLUP method. However, in comparison with the traditional ssGBLUP, the WssGBLUP increased the number of analyses steps and the gain obtained in predictive ability may be negligible. Moreover, the ssGBLUP resulted in the best predictive abilities and biases for androstenone using the ML population.

On the other hand, in general, the SL population result in better predictive ability in genomic prediction for indole and skatole.

**Keywords:** Androstenone. Gene interactions. Indole. Reference population. Skatole. Weighted genomic prediction.

### 3.2.Introduction

Boar taint, an unpleasant taste and smell of pork, is detected especially at cooking and it is caused by the increase of androstenone and skatole levels in adipose tissue from non-castrated male pigs. All pigs present skatole production in the large intestine. However, in non-castrated male pigs, the androstenone, which is produced in the testis, avoids skatole degradation (Zamaratskaia and Squires, 2009). Consequently, non-castrated male pigs have greater skatole deposition than barrows and gilts. In addition, other components, such as the indoles (4-phenyl-3-butenone, p-cresol and 4-ethylpheno) may also affect boar taint (Aluwé et al., 2011; Rius et al., 2005). Similar to skatole, indole is also synthesized from the tryptophan in the hindgut (Rius et al., 2005).

It is already known that androstenone, skatole and indole concentrations in adipose tissue show great genetic variability and depend on several factors as diet, age and genetic background (Aluwé et al., 2011; Campos et al., 2015; Duijvesteijn et al., 2010; Mathur et al., 2014). In addition, positive and favorable genetic correlations between skatole and androstenone levels have been described (Campos et al., 2015; Mathur et al., 2014). The variation of androstenone and skatole levels has been associated with single nucleotide polymorphism (SNP) markers and some candidate genes for boar taint have been reported (Campos et al., 2015; Duijvesteijn et al., 2014, 2010).

Besides being useful to detect potential marker candidates for assisted selection strategies, genome-wide association studies (GWAS) results can also contribute to improve the predictitive ability of genomic predictions. Several studies have proposed strategies to increase the accuracy and predictitive ability of genomic prediction by weighting SNP in the genomic relationship (**G**) matrix according to its relevance for the evaluated trait (Gao et al., 2017; Sarup et al., 2016; Veroneze et al., 2016; Wang et al., 2014, 2012; Zhang et al., 2016, 2010). In this approach, an iterative procedure is used to build a weighted **G** matrix based on the explained genetic variance by each marker, which can be used in a weighted single-step GBLUP (WssGBLUP) (Wang et al., 2012).

Despite the increase in the accuracy observed in some studies with WssGBLUP (Wang et al., 2012; Zhang et al., 2016), weights may be attributed to false positive markers, since there is no verification of the association. A way to increase emphasis on non-false positive markers could be to explore biological information, genetic correlations between traits and gene networks through an Association Weighted Matrix (AWM) (Fortes et al., 2010). The AWM applies gene network theory in GWAS to improve the identification of important set of genes explaining a group of phenotypes, which may be useful for deriving SNP weights to be posteriorly used to build the weighted **G** matrix. In brief, AWM methodology involves selecting SNPs from GWAS that are associated with genes that potentially explain a key phenotype. The SNP effects for *n* traits are used to build the AWM, which have as many rows as SNPs selected and as many columns as the number of evaluated traits. Each cell value of AWM corresponds to the normalized additive SNP effect on the trait. The rows are indexed as genes linked to SNPs and row-wise Pearson partial correlations are used to explore the correlations between SNP effects and to predict gene interactions using a combination of hierarchical clustering, weighted gene network, and pathway analyses (Fortes et al., 2011, 2010).

Despite produce relevant results in GWAS, contributing for the identification of candidate genes for complex traits, the incorporation of biological information from AWM into genomic prediction still lacks understanding. Therefore, we aimed to evaluate the predictive ability and the bias achieved using models that allow weighting each SNP in the **G** matrix according to its explained genetic variance, with and without incorporating biological information in the weighting procedure. In addition, we aimed to compare the predictive ability and the bias obtained from weighting approaches with those achieved using the traditional single-step GBLUP (ssGBLUP). Finally, we also aimed to evaluate the effect of different reference population structures (multi-line and single-line) in the predictive ability and the bias of boar taint compounds.

### **3.3. Materials and methods**

The data used for this study were obtained as part of routine data recording in a commercial breeding program. Samples collected for DNA extraction were only used for routine diagnostic purpose of the breeding program. Data recording and sample collection were conducted strictly in line with the rules given by Dutch Animal Research Authorities.

## **Data**

Genotypic and phenotypic data of animals from three sire lines (L1: Duroc-based line; L2: synthetic line; L3: Pietrain) were evaluated, as described in Table 1. The phenotypic information consisted of androstenone, skatole and indole levels measured in the adipose tissue of the carcass of 4,922 pigs as described in Mathur et al. (2014). Briefly, androstenone concentration was determined using liquid chromatography-mass spectrometry (Verheyden et al., 2007), whereas indole and skatole contents were measured using fluorescence at 285 and 340 nm (Ampuero Kragten et al., 2011).

Androstenone, skatole, and indole levels were submitted to logarithmic transformation (log), since the variables approximately followed log-normal distributions (Duijvesteijn et al., 2010; Mathur et al., 2014). Genotypic information of 3,749 animals from the three evaluated lines was also available for this study. The total number of animals in the pedigree was 13,604.

Table 1. Description of the number of animals with phenotypic and/or genotypic data from three sire lines.

| Line <sup>1</sup> | Animals    |           |                        |
|-------------------|------------|-----------|------------------------|
|                   | Phenotyped | Genotyped | Phenotyped + Genotyped |
| L1                | 3,572      | 1,316     | 854                    |
| L2                | 712        | 1,080     | 232                    |
| L3                | 638        | 1,353     | 123                    |
| Total             | 4,922      | 3,749     | 1,209                  |

<sup>1</sup> Duroc-based line (L1), Synthetic line (L2) and Pietrain (L3).

Two different populations were used in the analyses: a multi-line (ML) population with all three sire lines (L1, L2 and L3) and a single-line (SL) population composed by sire line L1, since it was the line with the greatest amount of genotyped and phenotyped animals.

## **Genotypes and Quality Control**

The animals were genotyped using the Illumina PorcineSNP60 BeadChip. The genotypic data were submitted to quality control, in which we excluded SNPs located in both sex chromosomes, with call-rate smaller than 95%, MAF smaller than 1% and/or with strong deviations from the Hardy-Weinberg equilibrium ( $P < 10^{-7}$ ). Quality control was performed within line, resulting in a final set of 49,977 SNPs for L1, 48,396 for L2 and 50,456 for L3. After quality control, the remaining missing genotypes were imputed within population using Fimpute v2.2 (Sargolzaei et al., 2014).

## Models

The analyses were conducted according to the following single trait mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}$$

Wherein:

$\mathbf{y}$  is a vector containing the logarithm of androstenone, skatole or indole levels;

$\mathbf{X}$  is the incidence matrix of fixed effects;

$\boldsymbol{\beta}$  is a vector of fixed effects, containing the effects of contemporary group (farm-year-month of slaughter), the covariates age at slaughter and scaled hot carcass weight at slaughter in each line and, additionally, the line effect in the ML population;

$\mathbf{Z}$  is the incidence matrix of animal additive genetic effect;

$\mathbf{a}$  is a vector of animal additive genetic effect,  $\mathbf{a} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{H})$ ;

$\mathbf{e}$  is a vector of residual effects,  $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$ ;

$\sigma_a^2$  and  $\sigma_e^2$  are the additive genetic and residual variances, respectively;

$\mathbf{I}$  is an identity matrix;

$\mathbf{H}$  is the relationship matrix based on both pedigree and genomic information, which inverse ( $\mathbf{H}^{-1}$ ) was given by Legarra et al. (2014):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

Wherein:

$\mathbf{A}^{-1}$  is the inverse of pedigree-based relationship matrix ( $\mathbf{A}$ );

$\mathbf{G}^{-1}$  is the inverse of the genomic relationship matrix;

$\mathbf{A}_{22}^{-1}$  is the inverse of pedigree-based relationship matrix from genotyped animals.

The  $\mathbf{G}$  matrix was calculated according to VanRaden (VanRaden, 2008; Wang et al., 2012):

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{D}\mathbf{Z}'}{2 \sum p_i q_i}$$

Wherein:

$\mathbf{Z}$  is a zero-centered matrix obtained by  $\mathbf{Z} = \mathbf{M} - \mathbf{P}$ , wherein  $\mathbf{M}$  is a  $m \times n$  (number of markers x number of animals) matrix, which specifies each individual genotype and  $\mathbf{P}$  is a matrix with the allele frequencies expressed as a difference of 0.5 and multiplied by 2, i.e. the  $i$  column of  $\mathbf{P}$  is given by  $2(p_i - 0.5)$ ;

$\mathbf{D}$  is a diagonal matrix, which will be better defined below;

$p_i$  and  $q_i$  are the SNP allelic frequencies in  $i^{\text{th}}$  loci.

### ***Scenarios and Weighted Matrix***

The first scenario (S1) is the traditional ssGBLUP (VanRaden, 2008), which will be taken as a reference, since it is the most widely used method in animal genomic prediction. In this approach, the **D** matrix corresponds to an identity matrix in which all markers are equally important in the construction of the relationship matrix.

The second scenario (S2) is the WssGBLUP method described by Wang et al. (2012). In this method, the breeding values obtained through the ssGBLUP are used to calculate SNP effects, which in turn are applied in the computation of the variance explained by each marker; then, these variances are used to build the **D** matrix. We considered the results from the third iteration of WssGBLUP, since it has been shown that three iterations are enough to maximize genomic predictive ability and correctly identify major SNPs (Lourenco et al., 2017; Zhang et al., 2016).

Aiming to incorporate biological information in the weighting matrix (**D**), we have proposed the AWM-WssGBLUP method (third scenario - S3). In this scenario, an adaptation of the approach presented by Fortes et al. (2010) was carried out to build the AWM and gene networks. In Fortes et al. (2010), a traditional single-SNP GWAS was performed and *P*-values were obtained for each SNP effect; then, significant SNP effects were used to build the AWM. In the present study, the AWM was built using the explained genetic variance by each SNP for each boar taint trait (androstenone, skatole, and indole). We have chosen to use the variance explained by each SNP since in a single-step GWAS, performed as back-solving the breeding values from ssGBLUP, SNP effects are computed using data from genotyped and non-genotyped animals. The AWM was built in four sub-scenarios, in which we selected the top 1%, 2%, 5% or 10% SNPs (top SNPs) that explained the highest proportion of the genetic variance for each trait ( $S3_{\text{top1\%}}$ ,  $S3_{\text{top2\%}}$ ,  $S3_{\text{top5\%}}$ , and  $S3_{\text{top10\%}}$ , respectively). All procedures adapted to build the AWM are summarized in Fig. 1 using the top 2% SNPs as an example.

For each sub-scenario, all markers identified as top SNPs for each trait were selected for AWM construction, following the methodology described by Fortes et al. (2010). Androstenone was considered the key phenotype in the analyses, since its level in non-castrated pigs directly affects skatole and indole depositions in adipose tissue (Moe et al., 2009; Rowe et al., 2014). In this way, SNPs identified as top SNPs for androstenone and for one of the other traits (skatole or indole) were selected to be used in the AWM.



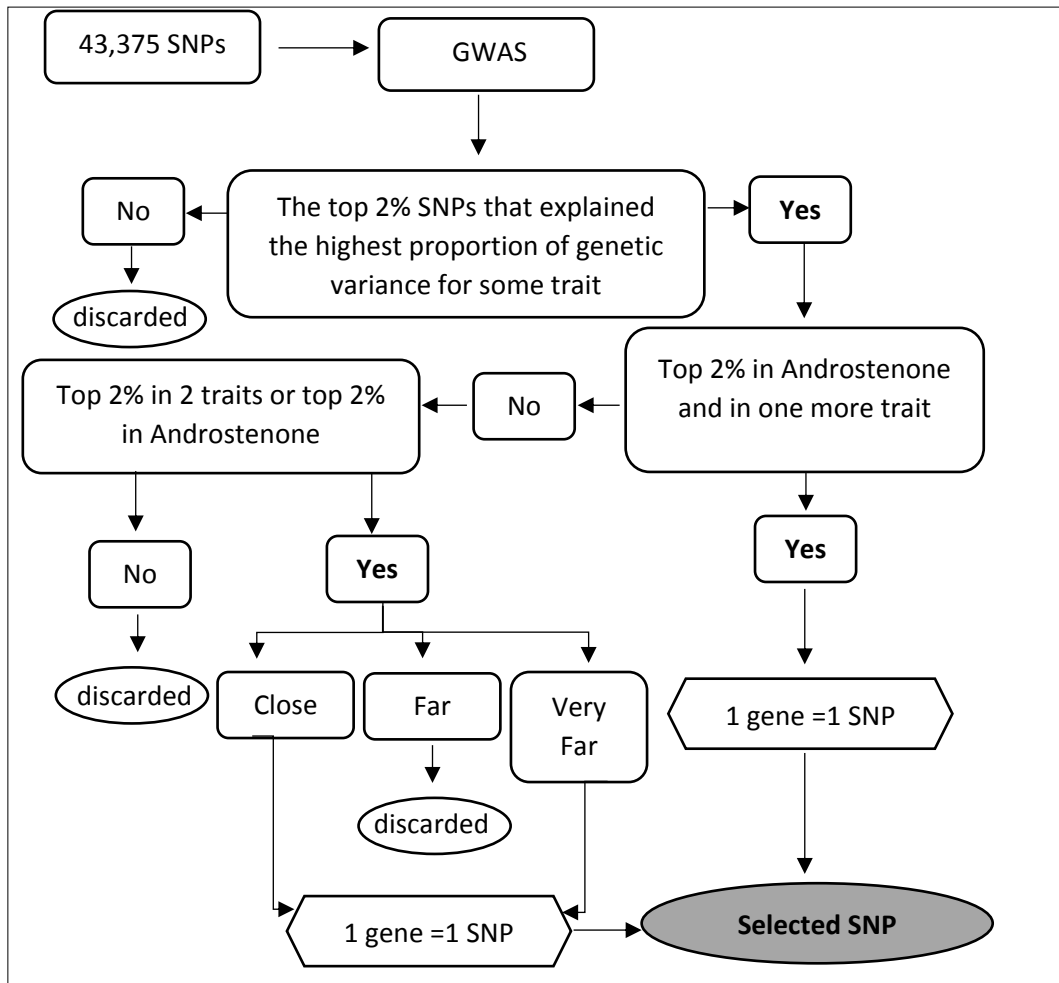


Figure 1: Scheme demonstrating the procedure for SNPs selection to build the Association Weight Matrix (AWM).

In order to perform the gene network analyses, we selected the genes closest to each top SNP using the Map2NCBI package (Hanna and Riley, 2014) of R software (R Core Team, 2017). The SNPs were classified as "close" (<2.5 kb), "far" ( $\geq 2.5$  kb and <400 kb) or "very far" ( $\geq 400$  kb) according to the distance of the nearest gene. Definitions for close or very far SNPs were based on linkage disequilibrium (LD) estimates in commercial pig lines (Veroneze et al., 2014), which showed that in several pig populations, including the sire lines used in this study, the LD ranges from 0.18 to 0.30 at distances from 200 to 500 kb. SNPs classified as far from annotated genes were discarded, since they were located substantially far from any coding region and this distance is not enough to justify inter loci interactions, as for very far SNPS (Fortes et al., 2010).

When more than one SNP was related to the same gene, we considered that this gene represented the SNP that met the following criteria: (1) was top in a higher number of traits, (2) showed the highest variance explained or (3) was the closest to the gene. After these steps, a final set of SNPs was selected and used to build the AWM with as many rows as the identified genes and as many columns as traits under study (three). Each cell  $\{i, j\}$  in AWM was completed with normalized (z-score) allelic substitution effect of  $i^{\text{th}}$  SNP in trait  $j$ .

Partial Pearson correlations between columns (three traits) and between rows (selected genes) of the AWM were calculated using unsupervised hierarchical clustering (Suppl. Fig. 1 and 2) in Hierarchical Clustering Explorer 3.5 software (HCE 3.5) (Jinwook Seo and Shneiderman, 2002). Significant correlations among AWM rows were identified with the PCIT algorithm proposed by Reverter and Chan (2008). The algorithm attributes a zero value for non-significant correlations and keeps the computed partial correlation for the significant ones. Thus, correlations were used as indicators of gene interactions (edges) in gene network analyses, which were performed in Cytoscape software (Shannon et al., 2003) (Suppl. Fig. 3 and 4). The total number of edges for each gene in the network was computed to identify which were the most important genes in the network. The number of edges was normalized (z-score) and the absolute values were used to build the **D** matrix previously described. Weight zero was given to SNPs that were not included in AWM.

The number of selected SNPs, identified genes and genes used to build the gene networks for each sub-scenario are shown in Table 2.

Table 2: Number of SNPs and genes used to build the Association Weight Matrix (AWM) and the gene networks.

| Scenarios <sup>1</sup> | Selected SNPs   |                 | Identified genes |       | Genes in the network |     |
|------------------------|-----------------|-----------------|------------------|-------|----------------------|-----|
|                        | SL <sup>2</sup> | ML <sup>3</sup> | SL               | ML    | SL                   | ML  |
| S3 <sub>Top1%</sub>    | 489             | 433             | 233              | 409   | 180                  | 122 |
| S3 <sub>Top2%</sub>    | 979             | 867             | 452              | 795   | 293                  | 163 |
| S3 <sub>Top5%</sub>    | 2,447           | 2,167           | 1,073            | 1,764 | 570                  | 375 |
| S3 <sub>Top10%</sub>   | 4,895           | 4,337           | 1,992            | 2,988 | 973                  | 538 |

<sup>1</sup> Sub-scenarios of AWM-WssGBLUP method with 1% (S3<sub>Top1%</sub>), 2% (S3<sub>Top2%</sub>), 5% (S3<sub>Top5%</sub>) and 10% (S3<sub>Top10%</sub>) of SNPs that explained the highest proportion of genetic variance.

<sup>2</sup> Single-line;

<sup>3</sup> Multi-line;

Additionally, a fourth scenario (S4) considering the number of edges obtained in S3<sub>top2%</sub> was built. The S3<sub>top2%</sub> was chosen since it presented the most consistent predictive ability for

all traits. In this scenario, each SNP was weighted by the normalized number of edges, considering the zero during the normalization. In other words, SNPs not included in AWM received the same weight, but it was different from zero.

The **D** matrix was used to compute six **G** matrices (one for S2, one for each S3 sub-scenario and one for S4). Prediction of genomic breeding values for all scenarios was performed using the BLUPF90 family of programs (Misztal et al., 2015). The proportions of variance explained by each marker in all scenarios are presented in Suppl. Fig. 5 to 10.

### ***Training and Validation Populations***

Aiming to evaluate the effect of different reference population size and composition in the predictive ability and the bias under all scenarios, marker effects were estimated considering two reference populations: a multi-line (ML) population with all three sire lines (L1, L2 and L3) and a single-line (SL) population composed by sire line L1. The 215 youngest animals from L1 with phenotypic and genotypic information were used as the validation population, since it was the line with the greatest amount of genotyped and phenotyped animals.

The prediction ability was computed as the Pearson correlation between the genomic estimated breeding value (GEBV) and the logarithm of phenotypes corrected for fixed effects. The prediction bias was computed as the difference between the unit and the linear regression coefficient of the phenotype logarithm corrected for fixed effects on the predicted GEBVs in each scenario.

## **3.4.Results**

### ***Variance components and genetic parameters***

Variance components, heritabilities and standard errors of SL and ML populations are presented in Table 3. Heritabilities ranging from 0.27 to 0.56 were estimated for all traits using both reference population scenarios.

Table 3. Variance components and heritabilities estimates for androstenone, skatole, and indole estimated using the single (SL) and the multi-line (ML) reference population.

| Populations <sup>1</sup> | Boar taint compounds | $\sigma_a^2$ (SE) <sup>2</sup> | $\sigma_e^2$ (SE) <sup>3</sup> | $h^2$ (SE) <sup>4</sup> |
|--------------------------|----------------------|--------------------------------|--------------------------------|-------------------------|
| SL                       | Androstenone         | 0.48 (0.07)                    | 0.79 (0.05)                    | 0.38 (0.04)             |
|                          | Skatole              | 0.13 (0.02)                    | 0.26 (0.01)                    | 0.33 (0.04)             |
|                          | Indole               | 0.07 (0.01)                    | 0.20 (0.01)                    | 0.27 (0.04)             |
| ML                       | Androstenone         | 0.69 (0.07)                    | 0.54 (0.04)                    | 0.56 (0.04)             |
|                          | Skatole              | 0.25 (0.02)                    | 0.24 (0.01)                    | 0.51 (0.03)             |
|                          | Indole               | 0.13 (0.01)                    | 0.20 (0.09)                    | 0.39 (0.03)             |

<sup>1</sup> Reference population composed by a Single-line (SL) or by Multi-line (ML);

<sup>2</sup> Additive genetic variance and standard error;

<sup>3</sup>  $\sigma_e^2$ : residual variance and standard error;

<sup>4</sup>  $h^2$ : heritability and standard error.

Moreover, when evaluating the effects of line in the ML population, we observed that L3 was the line with the lowest levels of all boar taint compounds, whereas L2 showed the greatest effects on androstenone and skatole levels and L1 presented the greatest effect on indole levels (data not shown).

### ***Evaluation of Methods***

The AWM-WssGBLUP, ssGBLUP, and WssGBLUP methods were evaluated in SL and ML populations. The methods were compared using the predictive ability (Table 4) and bias (Table 5).

Table 4: Predictive ability for boar taint compounds.

| Populations <sup>1</sup> | Scenarios <sup>2</sup> | Androstenone <sup>3</sup> | Skatole <sup>4</sup> | Indole <sup>5</sup> |
|--------------------------|------------------------|---------------------------|----------------------|---------------------|
| SL                       | S1                     | 0.314                     | 0.472                | 0.243               |
|                          | S2                     | 0.273                     | 0.432                | 0.238               |
|                          | S3 <sub>Top1%</sub>    | 0.298                     | 0.471                | 0.238               |
|                          | S3 <sub>Top2%</sub>    | 0.297                     | 0.471                | 0.255               |
|                          | S3 <sub>Top5%</sub>    | 0.283                     | 0.492                | 0.255               |
|                          | S3 <sub>Top10%</sub>   | 0.272                     | 0.462                | 0.251               |
|                          | S4                     | 0.297                     | 0.461                | 0.236               |
| ML                       | S1                     | 0.358                     | 0.467                | 0.229               |
|                          | S2                     | 0.322                     | 0.423                | 0.235               |
|                          | S3 <sub>Top1%</sub>    | 0.311                     | 0.469                | 0.240               |
|                          | S3 <sub>Top2%</sub>    | 0.309                     | 0.468                | 0.244               |
|                          | S3 <sub>Top5%</sub>    | 0.286                     | 0.461                | 0.238               |
|                          | S3 <sub>Top10%</sub>   | 0.284                     | 0.469                | 0.221               |
|                          | S4                     | 0.346                     | 0.409                | 0.210               |

<sup>1</sup> Reference population composed by a Single-line (SL) or by Multi-line (ML);

<sup>2</sup> Scenarios of genomic prediction. S1: ssGBLUP; S2: WssGBLUP; S3<sub>Top1%</sub>, S3<sub>Top2%</sub>, S3<sub>Top5%</sub>, S3<sub>Top10%</sub>: sub-scenarios of S3 (AWM-WssGBLUP method) with 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of genetic variance, respectively. S4: AWM-WssGBLUP method considering zero edges in the normalization;

<sup>3</sup> predictive ability for androstenone genomic prediction;

<sup>4</sup> predictive ability for skatole genomic prediction;

<sup>5</sup> predictive ability for indole genomic prediction.

In ML and SL, ssGBLUP (S1) was the method that provided the best predictive ability for androstenone. For skatole and indole, the predictive capacities were similar among scenarios. In SL, the AWM-WssGBLUP (S3) slightly increased the predictive capacity by up to 4% when using the top 5% SNPs. The S3<sub>Top5%</sub> also presented the best predictive ability for indole and skatole in SL population. In the scenario S4, which the weighting matrix include all markers, we observed that for SL the predictive abilities were similar to the scenarios that included just markers presented in AWM matrix. While, for ML population the predictive ability for androstenone were higher than in S3 scenarios and for skatole and indole the predictive ability in S4 was lower than in S3.

Table 5: Bias of prediction for boar taint compounds.

| Populations <sup>1</sup> | Scenarios <sup>2</sup> | Androstenone <sup>3</sup> | Skatole <sup>4</sup> | Indole <sup>5</sup> |
|--------------------------|------------------------|---------------------------|----------------------|---------------------|
| SL                       | S1                     | 0.434* <sup>6</sup>       | -0.236*              | 0.270*              |
|                          | S2                     | 0.631*                    | 0.170*               | 0.526*              |
|                          | S3 <sub>Top1%</sub>    | 0.393*                    | -0.463*              | 0.215*              |
|                          | S3 <sub>Top2%</sub>    | 0.401*                    | -0.460*              | 0.157*              |
|                          | S3 <sub>Top5%</sub>    | 0.427*                    | -0.498*              | 0.163*              |
|                          | S3 <sub>Top10%</sub>   | 0.457*                    | -0.394*              | 0.215*              |
|                          | S4                     | 0.540*                    | -0.061*              | 0.445*              |
| ML                       | S1                     | 0.324*                    | -0.254*              | 0.273*              |
|                          | S2                     | 0.551*                    | 0.186*               | 0.505*              |
|                          | S3 <sub>Top1%</sub>    | 0.408*                    | -0.533*              | 0.106*              |
|                          | S3 <sub>Top2%</sub>    | 0.411*                    | -0.442*              | 0.114*              |
|                          | S3 <sub>Top5%</sub>    | 0.484*                    | -0.328*              | 0.204*              |
|                          | S3 <sub>Top10%</sub>   | 0.474*                    | -0.416*              | 0.261*              |
|                          | S4                     | 0.563*                    | 0.905*               | 0.535*              |

<sup>1</sup> Reference population composed by a Single-line (SL) or by Multi-line (ML);

<sup>2</sup> Scenarios of genomic prediction. S1: ssGBLUP; S2: WssGBLUP; S3<sub>Top1%</sub>, S3<sub>Top2%</sub>, S3<sub>Top5%</sub>, S3<sub>Top10%</sub>: sub-scenarios of S3 (AWM-WssGBLUP method) with 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of genetic variance, respectively. S4: AWM-WssGBLUP method considering zero edges in the normalization;

<sup>3</sup> bias of prediction for androstenone genomic prediction;

<sup>4</sup> bias of prediction for skatole genomic prediction;

<sup>5</sup> bias of prediction for indole genomic prediction.

<sup>6</sup> significantly non-zero ( $p < 0.01$ ).

The prediction bias presented great variation among scenarios. Prediction for androstenone presented lower biases in S1 and S3<sub>Top1%</sub> in ML and SL, respectively. For skatole, lower biases were observed in S2 and S4 in ML and SL, respectively. For indole, lower biases were verified for S3<sub>top1%</sub> and S3<sub>top2%</sub> in ML and SL, respectively. Moreover, comparing S3<sub>Top2%</sub> and S4, in both populations, the biases for androstenone and indole were lower in S3<sub>Top2%</sub> than in S4

### 3.5. Discussion

High heritabilities for androstenone, skatole and indole (0.56, 0.51 and 0.39, respectively) in the ML population and somewhat lower (0.38, 0.33 and 0.27, respectively) in SL were found in the present study. These differences on the heritabilities may be explained by population size and structure. Moreover, our findings are in agreement with previous studies (Campos et al.,

2015; Mathur et al., 2012; Windig et al., 2012), which reported heritabilities ranging from 0.46 to 0.72 for androstenone, 0.26 to 0.50 for skatole and 0.29 to 0.35 for indole.

It has been shown that genetic background affects the levels of boar taint compounds in pig carcass, since each breed presents specific deposition patterns of androstenone, skatole and indole (Gregersen et al., 2012; Grindflek et al., 2011). For example, Duroc breed presents greater deposition of androstenone than Landrace (Grindflek et al., 2011) and also greater deposition of skatole and indole than Yorkshire breed (Gregersen et al., 2012). We evaluated three different pig sire lines based on Duroc (L1), a synthetic line (L2) and Pietrain (L3) breed and observed that L1 and L2 had the greatest levels of all boar taint compounds.

In the current study, we proposed an adaptation of the AWM methodology described by Fortes et al. (2010) to find SNPs closely related to candidate genes for boar taint traits. Since androstenone level in non-castrated pigs directly affects the deposition of skatole and indole in adipose tissue, androstenone was chosen as a key trait to build the AWM. After the estimation of substitution allelic effects, we selected four groups of SNPs that correspond to the 1, 2, 5, or 10% SNPs that explained the highest proportion of genetic variance for each trait. These groups were used in the beginning of the process of building the AWM matrix. As expected, a higher initial number of SNPs resulted in a higher number of genes identified and used to build the gene networks. Moreover, we observed that when a SL population was used in GWAS, a higher number of genes were identified in comparison to ML population. Raven et al. (2014) demonstrated that multipopulation GWAS increases QTL mapping precision, since it reduces the LD between markers and QTL. However, in ML population, QTLs that are not segregating in all populations may be not identified, which could explain in part why a higher number of genes were identified in SL.

A high level of similarity between SNP effects for skatole and indole was observed, which may have occurred because skatole and indole are produced in the same metabolic pathway (Zamaratskaia and Squires, 2009). Moreover, there is a strong genetic correlation between skatole and indole, ranging from 0.71 to 0.78 (Grindflek et al., 2011; Lee et al., 2005). Although androstenone acts as an antagonist for skatole degradation (Aluwé et al., 2011; Doran et al., 2002), in the hierarchical clustering of SNP effects, this trait was located in a different branch in the cluster dendrogram. Different genetic mechanisms are probably involved on androstenone pathway and skatole and indole catabolism (Zamaratskaia and Squires, 2009), since genetic correlations between androstenone and skatole or indole have been described to be moderate, around 0.31 and 0.46 (Campos et al., 2015; Lee et al., 2005; Windig et al., 2012).

The ssGBLUP, as the traditional GBLUP, assumes that all markers equally contribute to the construction of the genomic relationship matrix (Goddard, 2009), ignoring any available information regarding the genetic architecture of the trait. However, it is well known that a finite number of genes control quantitative traits (Hayes and Goddard, 2001). Using models that allow highlighting SNPs related to candidate genes for a given trait may improve the results of genomic prediction compared to ssGBLUP (Wang et al., 2014).

GWAS have identified several QTL regions for boar taint traits (Campos et al., 2015; Duijvesteijn et al., 2015, 2014, 2010), however, there are no studies exploring GWAS results in the genomic prediction for boar taint. Using simulated data, Fragomeni et al (2017) showed that the inclusion of real quantitative trait nucleotide (QTN) effects as weights in the genomic relationship matrix, blending with 1% of identity matrix, resulted in an accuracy of 0.99 in genomic prediction. These authors stated that high accuracy could be obtained if realistic weights for the causative QTN are used. Thus, direct GWAS results may not be the best way to derive weights for WssGBLUP because of false positive associations. Combining GWAS and biological information may be a path for realistic weight estimation.

The AWM, proposed by Fortes et al. (2010), combines multiple traits and is used for building gene networks, which allow the identification of the most important regions in the trait control. The gene networks help to understand boar taint genetic architecture and consequently may provide better weights to be used in WssGBLUP. In our study, the normalized number of edges for each gene in the network was used to build the weighting matrix (**D**). In all S3 sub-scenarios, most of the genes presented only one edge; however, genes with four or more edges were also present, which resulted in different weights for the previously identified SNPs.

Based on the gene network analyses, we can infer that our findings are robust and consistent with the studied traits, since genes previously reported in literature controlling boar taint traits were identified. As an example, we can quote the genes *CYP2E1* (cytochrome P450IIE1), *PTPRT* (protein tyrosine phosphatase, receptor type T) and *HSD17B2* (estradiol 17-beta-dehydrogenase) that were reported in previous GWAS as being involved with boar taint appearance (Moe et al., 2009; Rowe et al., 2014). These genes identification shows that the AWM associated with gene networks pinpointed genome regions closely related to boar taint traits.

The predictive ability was higher using the traditional ssGBLUP than using the WssGBLUP or AWM-WssGBLUP. Moreover, WssGBLUP and AWM-WssGBLUP presented similar predictive capacity and bias in most evaluated scenarios. For androstenone, a reduction



in the predictive capacity was observed when a higher number of SNPs were selected after GWAS ( $S_{3_{top5\%}}$  and  $S_{3_{top10\%}}$ ). The better performance of traditional ssGBLUP compared to other methods may be explained by the polygenic genetic architecture of the evaluated traits, since the predictive ability of methods exploring weighted relationship matrices depends on the trait control and it is higher when the trait is controlled by only few QTLs (Fang et al., 2017; Gao et al., 2017; Lourenco et al., 2017; Veroneze et al., 2016; Zhang et al., 2010).

Reducing the number of SNPs used to build the  $\mathbf{G}$  matrix from more than 40,000 ( $S_1$ ) to less than 1,000 SNPs ( $S_{3_{Top10\%}}$ ) resulted in a loss of less than 0.06 points in model predictive capacity and small increases in prediction bias. This result shows the robustness of the genomic relationship, as previously shown by Lopes et al. (2013). Lenz et al. (2017) reported that in structured populations, even using few SNPs (500 to 1,000) spread across the genome to elaborate the  $\mathbf{G}$  matrix, small or no losses in predictitive ability are expected due to family structure and LD in the population. Up to 95% of accuracy obtained by using a higher density panel can be obtained by using only a small proportion of markers depending on the genetic architecture of the trait and the effective population size (Zhang et al., 2011).

Aiming to verify the effect of not excluding markers to build the genomic relationship matrix in  $S_3$ , we proposed the  $S_4$  scenario, wherein the SNPs not included in the top 2% SNPs received a non-zero weight. Our findings indicate that non-zero weighting, in most of the cases, resulted in decreased predictive abilities and increased biases for both SL and ML populations, indicating that SNPs with small effects may be excluded in weighted genomic prediction.

Assessing the effects of the reference population, we observed that the ML increased the prediction bias in most scenarios and traits. The ML reference population improved the predictive ability of androstenone, however, skatole and indole, in general, presented better predictive ability and bias in the SL reference population. It has been suggested that using multi- and larger populations in genomic selection increases the statistical power of the analyses (Stranger et al., 2011), since the lower LD, usually observed between markers in those populations, increases the ability to map the QTLs (Raven et al., 2014) especially when the traits present large SNP effects (Liu et al., 2011). Nevertheless, the prediction for all boar taint compounds would be less biased using SL reference population.

### **3.6. Conclusion**

In summary, using biological information, through AWM matrix and gene networks, to derive weights for genomic prediction resulted in slight increase in predictive ability for skatole

and indol. However, this approach increased the number of analyses steps. In addition, for androstenone the traditional ssGBLUP provided higher predictive ability in comparison to the weighted scenarios. Thus, we can conclude that ssGBLUP is most appropriate for the analysis of boar taint compounds in comparison to the weighted strategies used in the present work. In general, the single-line population result in better predictive ability in genomic prediction for indole and skatole.

### 3.7. References

- Aluwé, M., S. Millet, K.M. Bekaert, F.A.M.M. Tuytens, L. Vanhaecke, et al. 2011. Influence of breed and slaughter weight on boar taint prevalence in entire male pigs. *Animal* 5(8): 1283–1289. doi: 10.1017/S1751731111000164.
- Ampuero Kragten, S., B. Verkuylen, H. Dahlmans, M. Hortos, J.A. Garcia-Regueiro, et al. 2011. Inter-laboratory comparison of methods to measure androstenone in pork fat. *Animal* 5(10): 1634–1642. doi: 10.1017/S1751731111000553.
- Campos, C.F. de, M.S. Lopes, F.F. e Silva, R. Veroneze, E.F. Knol, et al. 2015. Genomic selection for boar taint compounds and carcass traits in a commercial pig population. *Livest Sci* 174: 10–17. doi: 10.1016/j.livsci.2015.01.018.
- Doran, E., F.W. Whittington, J.D. Wood, and J.D. Mcgivan. 2002. Cytochrome P450IIE1 ( CYP2E1 ) is induced by skatole and this induction is blocked by androstenone in isolated pig hepatocytes. *Chem Biol Interact* 140: 81–92. doi: 10.1016/S0009-2797(02)00015-7.
- Duijvesteijn, N., E.F. Knol, and P. Bijma. 2014. Boar taint in entire male pigs : A genomewide association study for direct and indirect genetic effects on androstenone. *J Anim Sci* 92: 4319–4328. doi: 10.2527/jas2014-7863.
- Duijvesteijn, N., E.F. Knol, and P. Bijma. 2015. Direct and associative effects for androstenone and genetic correlations with backfat and growth in entire male pigs. *J Anim Sci*: 2465–2475. doi: 10.2527/jas2011-4625.
- Duijvesteijn, N., E.F. Knol, J.W.M. Merks, R.P.M.A. Crooijmans, M.A.M. Groenen, et al. 2010. A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. *BMC Genet* 11(42): 1–11. doi: 10.1186/1471-2156-11-42.
- Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu, et al. 2017. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet Sel Evol* 49(1): 44. doi: 10.1186/s12711-017-0319-0.
- Fortes, M.R.S., A. Reverter, S.H. Nagaraj, Y. Zhang, N.N. Jonsson, et al. 2011. A single nucleotide polymorphism-derived regulatory gene network underlying puberty in 2 tropical breeds of beef cattle. *J Anim Sci* 89(6): 1669–1683. doi: 10.2527/jas.2010-3681.
- Fortes, M.R.S., A. Reverter, Y. Zhang, E. Collis, S.H. Nagaraj, et al. 2010. Association weight matrix for the genetic dissection of puberty in beef cattle. *Pnas* 107(31): 1–6. doi: 10.1073/pnas.1002044107.
- Fragomeni, B.O., D.A.L. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2017.

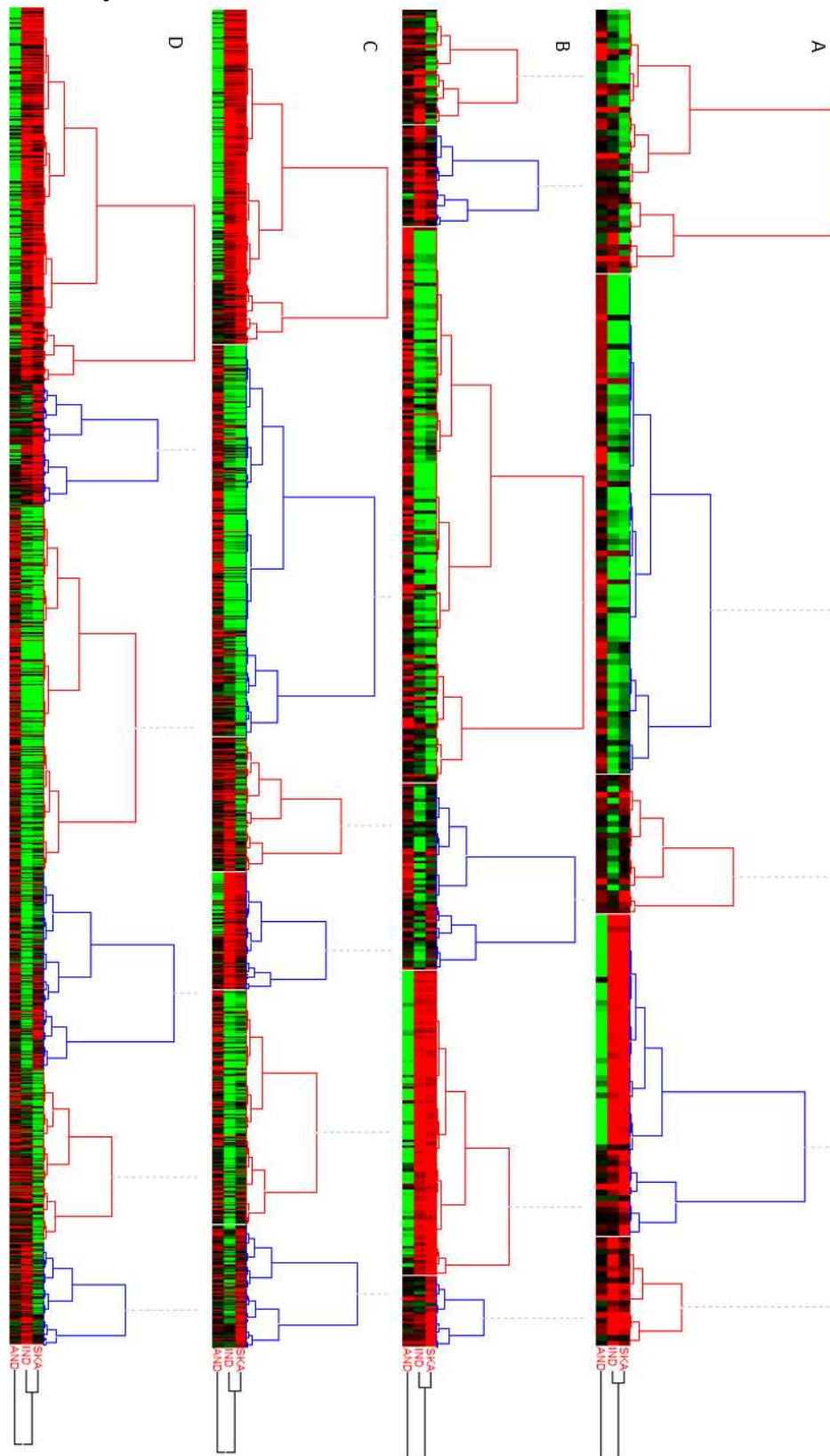
- Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genet Sel Evol*: 1–11. doi: 10.1186/s12711-017-0335-0.
- Gao, N., J.W.R. Martini, Z. Zhang, X. Yuan, H. Zhang, et al. 2017. Incorporating Gene Annotation into Genomic Prediction of Complex Phenotypes. *Genetics* 207(10): 489–501. doi: 10.1534/genetics.117.300198/-/DC1.1.
- Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136(2): 245–257. doi: 10.1007/s10709-008-9308-0.
- Gregersen, V.R., L.N. Conley, K.K. Sørensen, B. Guldbrandtsen, I.H. Velander, et al. 2012. Genome-wide association scan and phased haplotype construction for quantitative trait loci affecting boar taint in three pig breeds. *BMC Genomics* 13(1): 22. doi: 10.1186/1471-2164-13-22.
- Grindflek, E., T.H.E. Meuwissen, T. Aasmundstad, H. Hamland, M.H.S. Hansen, et al. 2011. Revealing genetic relationships between compounds affecting boar taint and reproduction in pigs. *J Anim Sci* 89(3): 680–692. doi: 10.2527/jas.2010-3290.
- Hanna, L.L.H., and D.G. Riley. 2014. Mapping genomic markers to closest feature using the R package Map2NCBI. *Livest Sci* 162: 59–65. doi: 10.1016/j.livsci.2014.01.019.
- Hayes, B.J., and M.E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet Sel Evol* 33(3): 209–229. doi: 10.1186/1297-9686-33-3-209.
- Jinwook Seo, and B. Shneiderman. 2002. Interactively exploring hierarchical clustering results [gene identification]. *Computer (Long Beach Calif)* 35(7): 80–86. doi: 10.1109/MC.2002.1016905.
- Lee, G.J., A.L. Archibald, A.S. Law, S. Lloyd, J. Wood, et al. 2005. Detection of quantitative trait loci for androstenone, skatole and boar taint in a cross between Large White and Meishan pigs. *Anim Genet* 36(1): 14–22. doi: 10.1111/j.1365-2052.2004.01214.x.
- Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. *Livest Sci* 166(1): 54–65. doi: 10.1016/j.livsci.2014.04.029.
- Lenz, P.R.N., J. Beaulieu, S.D. Mansfield, S. Clément, M. Despots, et al. 2017. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (*Picea mariana*). *BMC Genomics* 18(1): 1–17. doi: 10.1186/s12864-017-3715-5.
- Liu, Z., F.R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, et al. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet Sel Evol* 43(1): 19. doi: 10.1186/1297-9686-43-19.
- Lopes, M.S., F.F. Silva, B. Harlizius, N. Duijvesteijn, P.S. Lopes, et al. 2013. Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genet* 14(1): 92. doi: 10.1186/1471-2156-14-92.
- Lourenco, D.A.L., B.O. Fragomeni, H.L. Bradford, I.R. Menezes, J.B.S. Ferraz, et al. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. *J Anim Breed Genet* 134(6): 463–471. doi: 10.1111/jbgs.12288.
- Mathur, P.K., J. ten Napel, R.E. Crump, H.A. Mulder, and E.F. Knol. 2014. Genetic relationship between boar taint compounds, human nose scores, and reproduction traits in pigs. *J Anim Breed Genet* 91(9): 4080–4089. doi: <https://doi.org/10.2527/jas.2013->

6478.

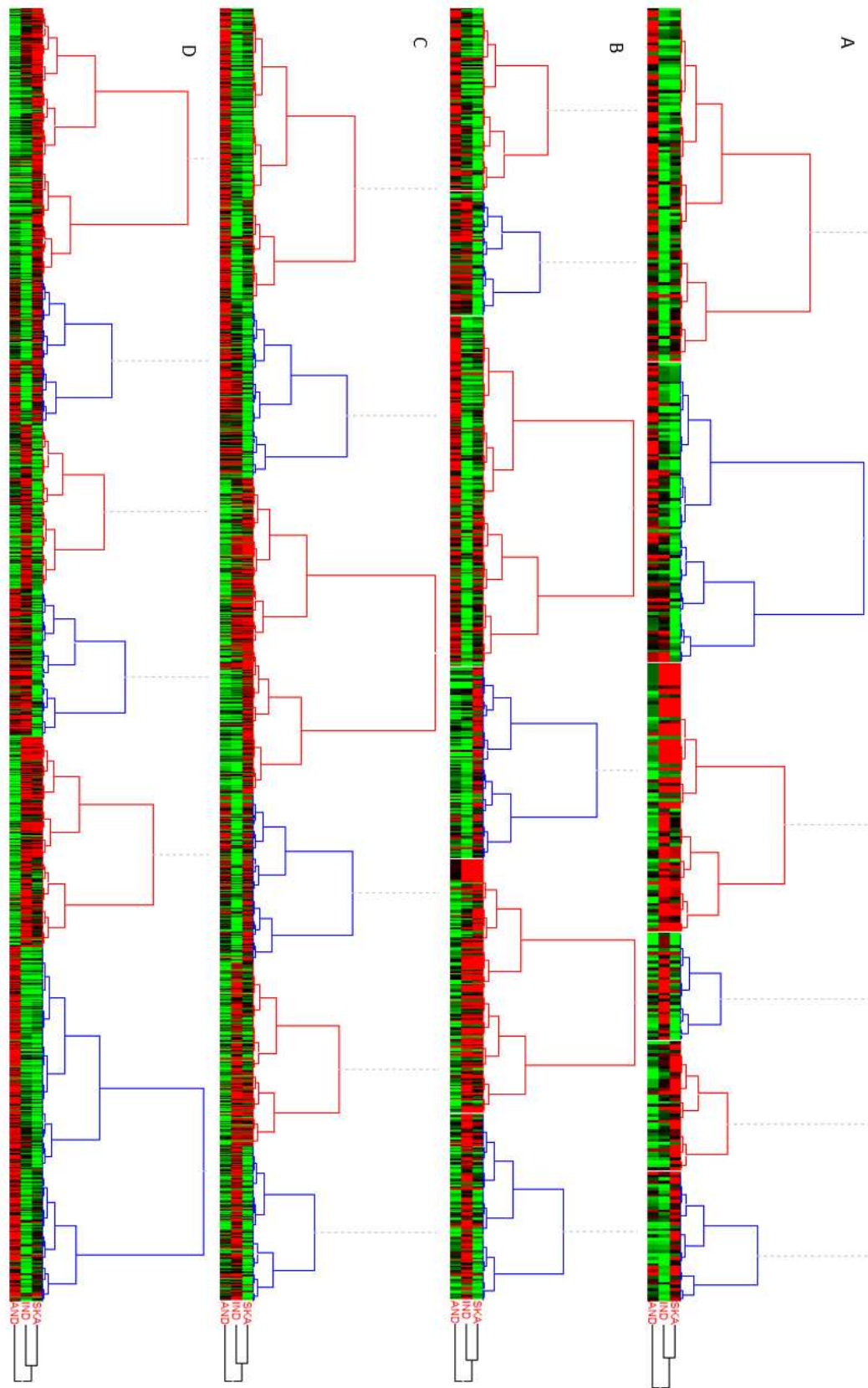
- Mathur, P.K., J. ten Napel, S. Bloemhof, L. Heres, E.F. Knol, et al. 2012. A human nose scoring system for boar taint and its relationship with androstenone and skatole. *Meat Sci* 91(4): 414–422. doi: 10.1016/j.meatsci.2012.02.025.
- Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, et al. 2015. Manual for BLUPF90 family of programs. Univ Georg Athens, USA: 125.  
[http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90\\_all2.pdf](http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf).
- Moe, M., S. Lien, T. Aasmundstad, T.H. Meuwissen, M.H. Hansen, et al. 2009. Association between SNPs within candidate genes and compounds related to boar taint and reproduction. *BMC Genet* 10(1): 32. doi: 10.1186/1471-2156-10-32.
- R Core Team. 2017. R: A language and environment for statistical computing.
- Raven, L., B.G. Cocks, and B.J. Hayes. 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics* 15(1): 62. doi: 10.1186/1471-2164-15-62.
- Reverter, A., and E.K.F. Chan. 2008. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24(21): 2491–2497. doi: 10.1093/bioinformatics/btn482.
- Rius, M.A., M. Hortós, and J.A. García-Regueiro. 2005. Influence of volatile compounds on the development of off-flavours in pig back fat samples classified with boar taint by a test panel. *Meat Sci* 71(4): 595–602. doi: 10.1016/j.meatsci.2005.03.014.
- Rowe, S.J., B. Karacaören, D.-J. de Koning, B. Lukic, N. Hastings-Clark, et al. 2014. Analysis of the genetics of boar taint reveals both single SNPs and regional effects. *BMC Genomics* 15(1): 424. doi: 10.1186/1471-2164-15-424.
- Sarup, P., J. Jensen, T. Ostensen, M. Henryon, and P. Sørensen. 2016. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet* 17(1): 11. doi: 10.1186/s12863-015-0322-9.
- Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, et al. 2003. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*: 2498–2504. doi: 10.1101/gr.1239303.metabolite.
- Stranger, B.E., E.A. Stahl, and T. Raj. 2011. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics* 187(2): 367–383. doi: 10.1534/genetics.110.120907.
- VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. *J Dairy Sci* 91(11): 4414–4423. doi: 10.3168/jds.2007-0980.
- Verheyden, K., H. Noppe, M. Aluwé, S. Millet, J. Vanden Bussche, et al. 2007. Development and validation of a method for simultaneous analysis of the boar taint compounds indole, skatole and androstenone in pig fat using liquid chromatography–multiple mass spectrometry. *J Chromatogr A* 1174(1–2): 132–137. doi: 10.1016/j.chroma.2007.08.075.
- Veroneze, R., J.W.M. Bastiaansen, E.F. Knol, S.E.F. Guimarães, F.F. Silva, et al. 2014. Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (*Sus scrofa*) populations. *BMC Genet* 15(1): 126. doi: 10.1186/s12863-014-0126-3.
- Veroneze, R., P.S. Lopes, M.S. Lopes, A.M. Hidalgo, S.E.F. Guimarães, et al. 2016.

- Accounting for genetic architecture in single- and multipopulation genomic prediction using weights from genomewide association studies in pigs. *J Anim Breed Genet* 133(3): 187–196. doi: 10.1111/jbg.12202.
- Wang, H., I. Misztal, I. Aguilar, A. Legarra, R.L. Fernando, et al. 2014. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Front Genet* 5(MAY): 1–10. doi: 10.3389/fgene.2014.00134.
- Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet Res (Camb)* 94(2): 73–83. doi: 10.1017/S0016672312000274.
- Windig, J.J., H.A. Mulder, J. ten Napel, E.F. Knol, P.K. Mathur, et al. 2012. Genetic parameters for androstenone, skatole, indole, and human nose scores as measures of boar taint and their relationship with finishing traits. *J Anim Sci* 90(7): 2120–2129. doi: 10.2527/jas.2011-4700.
- Zamaratskaia, G., and E.J. Squires. 2009. Biochemical, nutritional and genetic effects on boar taint in entire male pigs. *Animal* 3(11): 1508–1521. doi: 10.1017/S1751731108003674.
- Zhang, Z., X. Ding, J. Liu, Q. Zhang, and D.-J. de Koning. 2011. Accuracy of genomic prediction using low-density marker panels. *J Dairy Sci* 94(7): 3642–3650. doi: 10.3168/jds.2010-3917.
- Zhang, Z., J. Liu, X. Ding, P. Bijma, D.-J. de Koning, et al. 2010. Best Linear Unbiased Prediction of Genomic Breeding Values Using a Trait-Specific Marker-Derived Relationship Matrix (T. Mailund, editor). *PLoS One* 5(9): e12648. doi: 10.1371/journal.pone.0012648.
- Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. *Front Genet* 7(AUG): 1–14. doi: 10.3389/fgene.2016.00151.

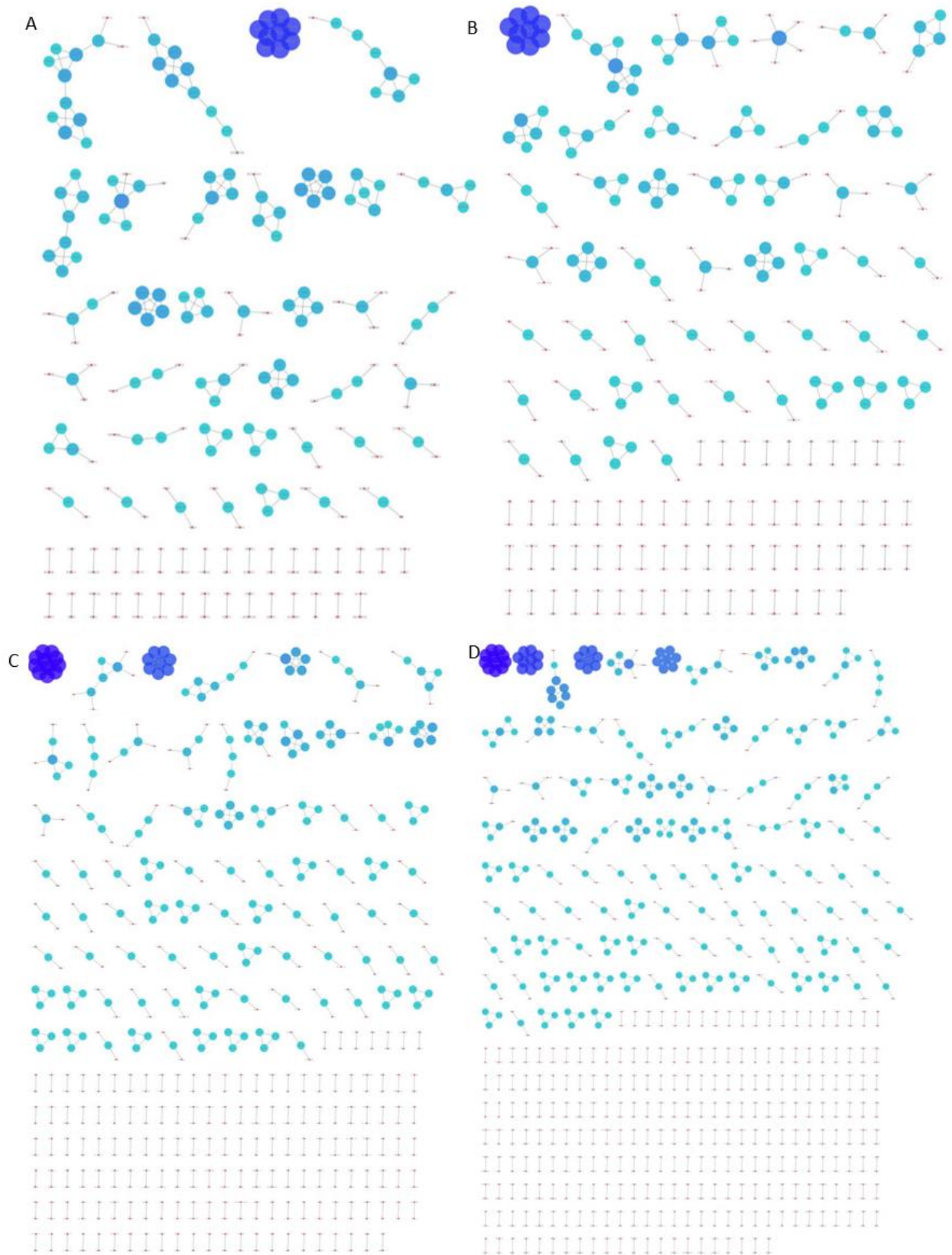
### 3.8. Supplementary material



Supplementary Figure 1: Unsupervised hierarchical clustering of SNP effect in a multi-line population. Sub-scenarios (A)  $S3_{\text{Top}1\%}$ , (B)  $S3_{\text{Top}2\%}$ , (C)  $S3_{\text{Top}5\%}$ , (D)  $S3_{\text{Top}10\%}$ , on which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively.

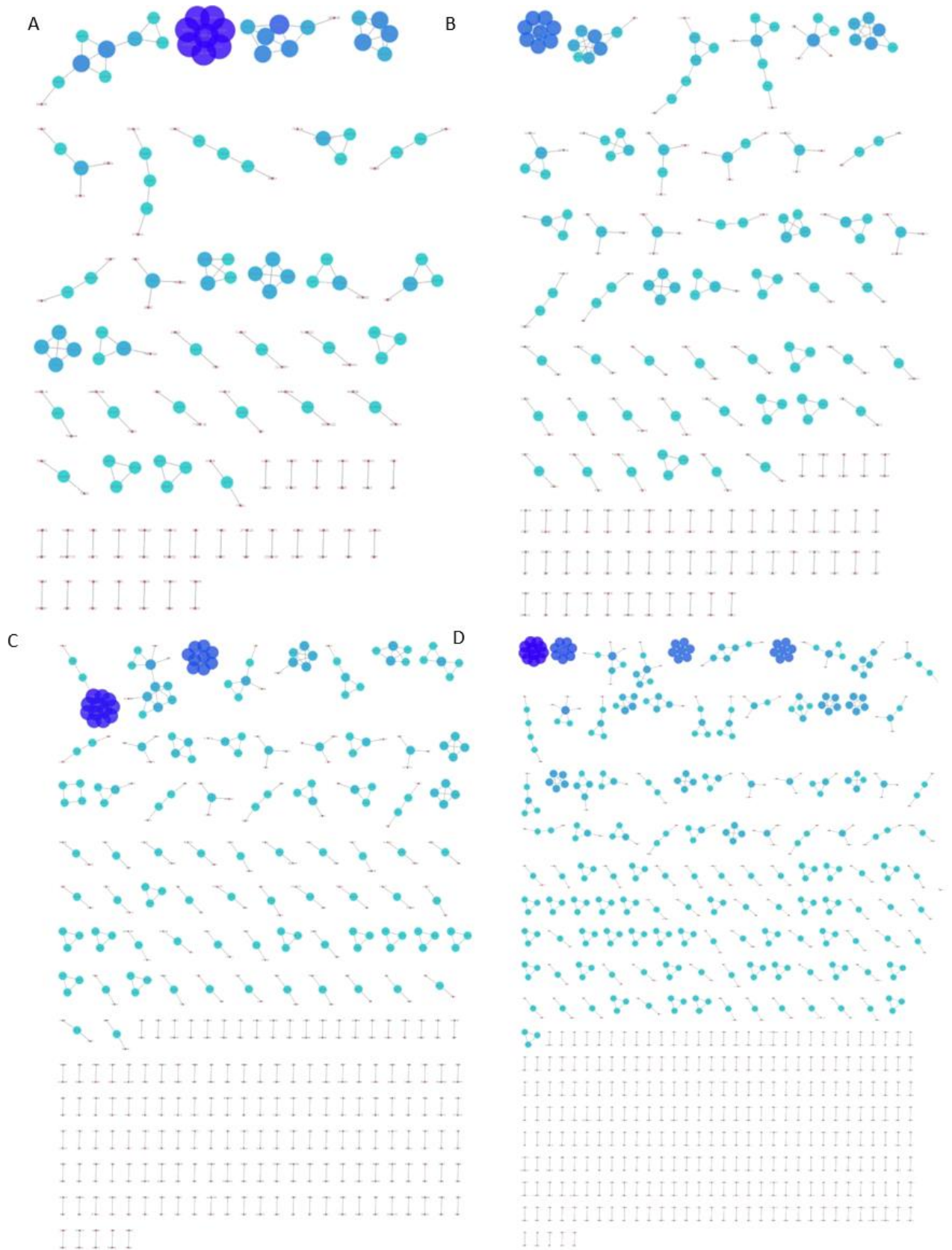


Supplementary Figure 2: Unsupervised hierarchical clustering of SNP effect in a single-line population. Sub-scenarios (A)  $S3_{\text{Top}1\%}$ , (B)  $S3_{\text{Top}2\%}$ , (C)  $S3_{\text{Top}5\%}$ , (D)  $S3_{\text{Top}10\%}$ , on which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively.

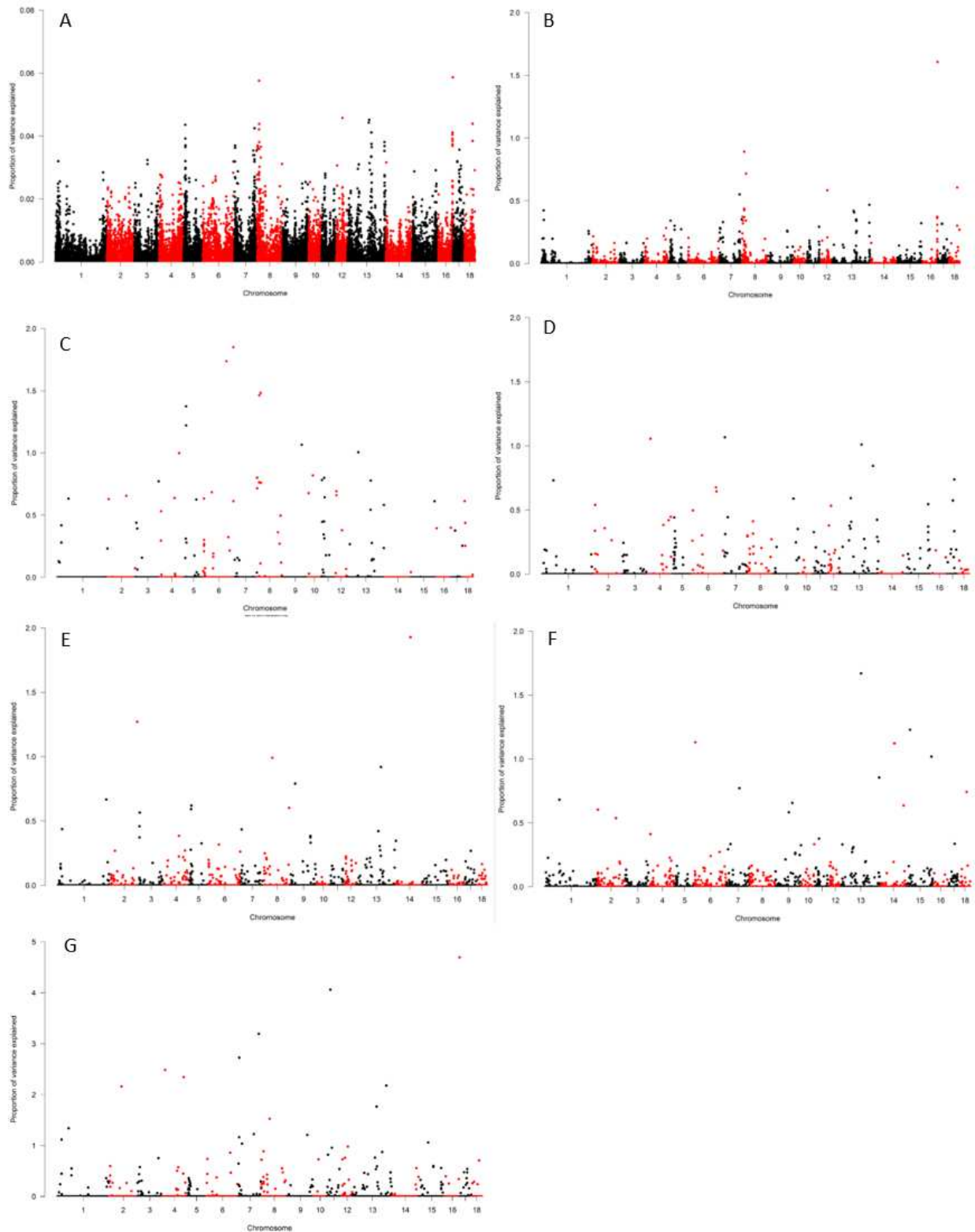


Supplementary Figure 3: Gene networks elaborated using genes identified in a multi-line population on sub-scenarios (A)  $S3_{Top1\%}$ , (B)  $S3_{Top2\%}$ , (C)  $S3_{Top5\%}$  and (D)  $S3_{Top10\%}$  based on an Association Weight Matrix (AWM), in which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively. Circles represent the genes and edges represent the interaction between adjacent genes.

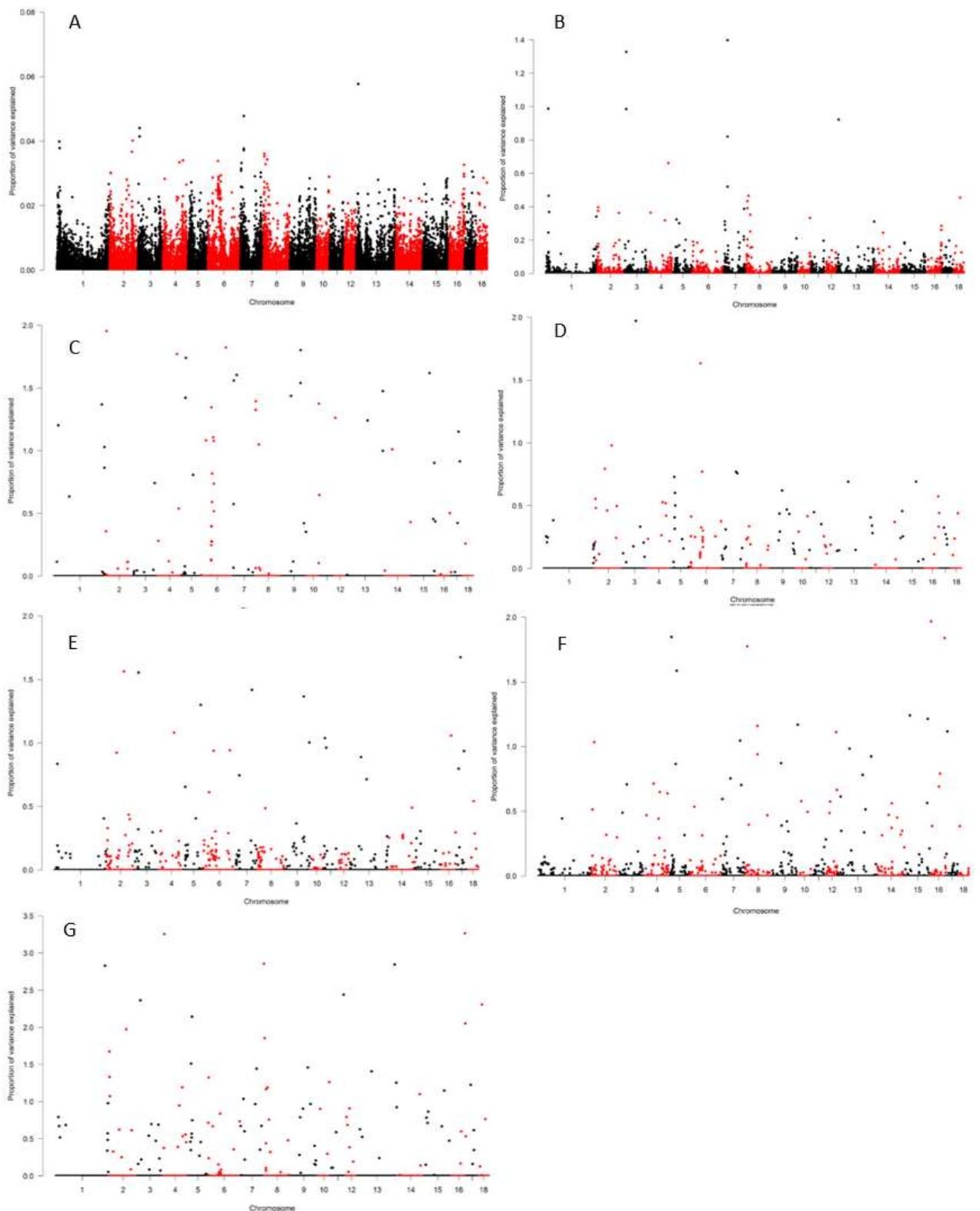




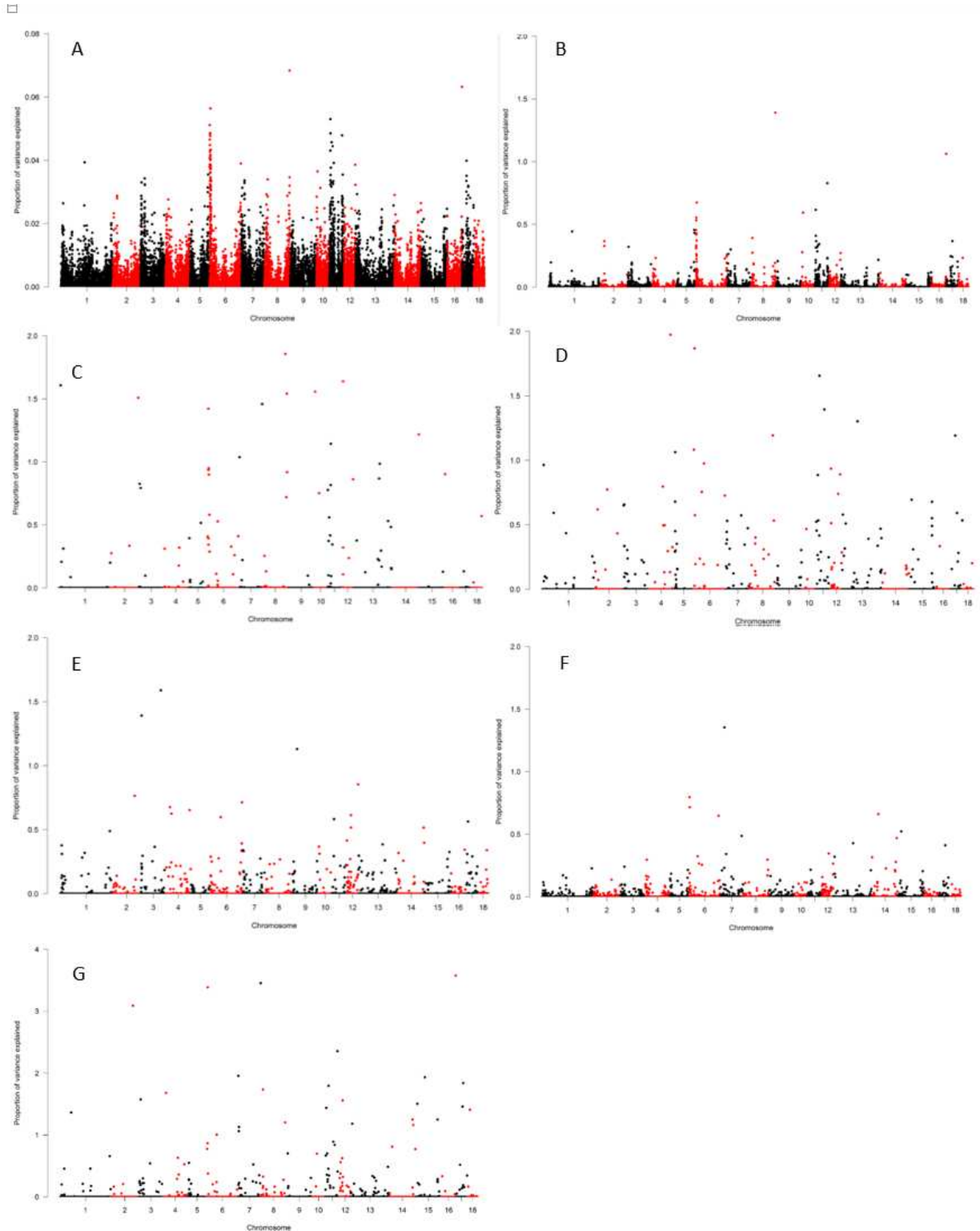
Supplementary Figure 4: Gene networks elaborated using genes identified in a single-line population on sub-scenarios (A)  $S3_{Top1\%}$ , (B)  $S3_{Top2\%}$ , (C)  $S3_{Top5\%}$  and (D)  $S3_{Top10\%}$ , based on an Association Weight Matrix (AWM), in which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively. Circles represent the genes and edges represent the interaction between adjacent genes.



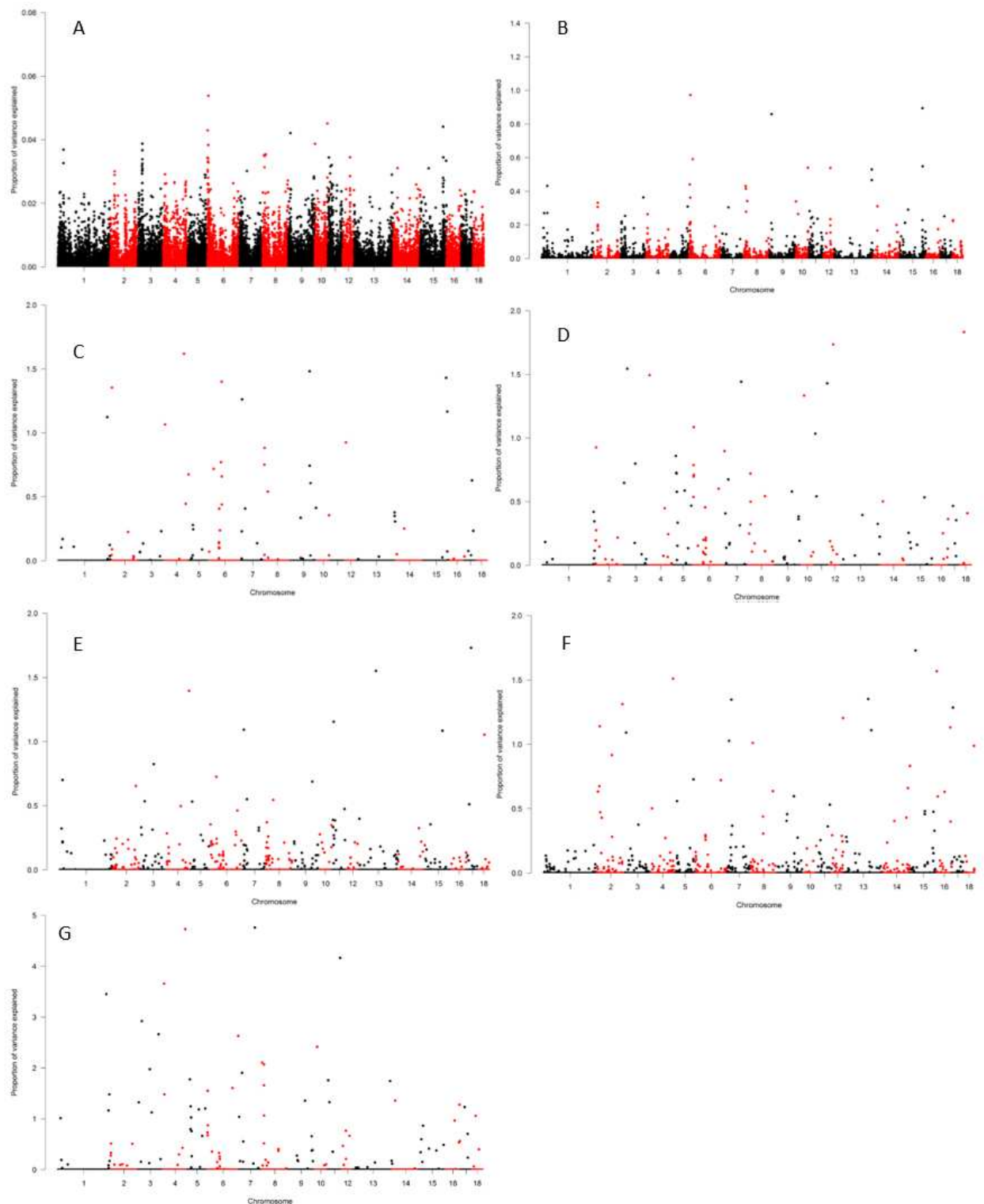
Supplementary Figure 5: Manhattan plots of proportion of explained genetic variance by SNPs for androstene levels in a single-line population in scenarios: (A) S1 (ssGBLUP); (B) S2 (WssGBLUP) and sub-scenarios based on AWM: (C) S3<sub>Top1%</sub>, (D) S3<sub>Top2%</sub>, (E) S3<sub>Top5%</sub>, (F) S3<sub>Top10%</sub>, in which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively; (G) S4 (AWM-WssGBLUP method considering zero edges in the normalization).



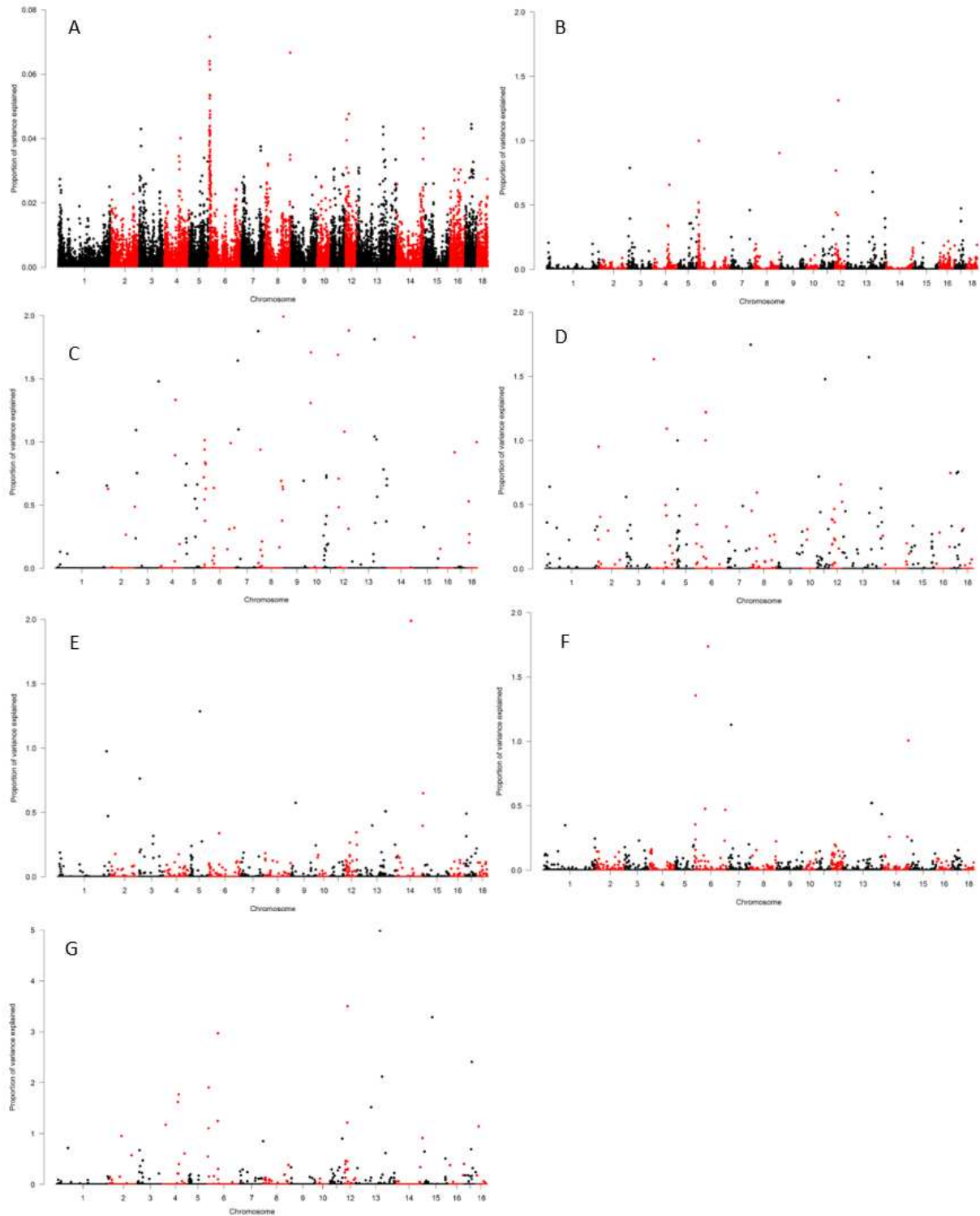
Supplementary Figure 6: Manhattan plots of proportion of explained genetic variance by SNPs for androstene levels in a multi-line population in scenarios: (A) S1 (ssGBLUP); (B) S2 (WssGBLUP) and sub-scenarios based on AWM: (C) S3<sub>Top1%</sub>, (D) S3<sub>Top2%</sub>, (E) S3<sub>Top5%</sub>, (F) S3<sub>Top10%</sub>, in which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively; (G) S4 (AWM-WssGBLUP method considering zero edges in the normalization).



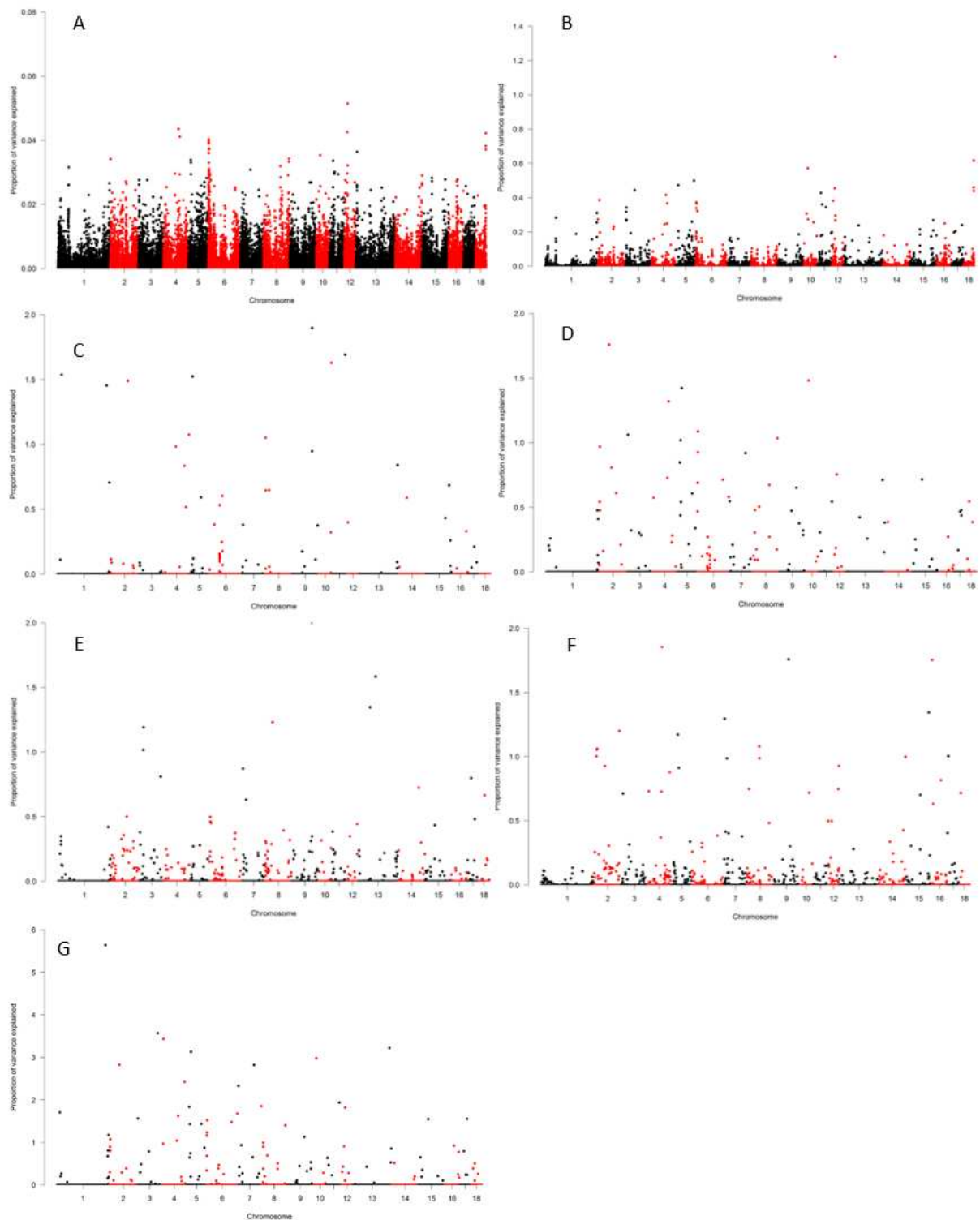
Supplementary Figure 7: Manhattan plots of proportion of explained genetic variance by SNPs for skatole levels in a single-line population in scenarios: (A) S1 (ssGBLUP); (B) S2 (WssGBLUP) and sub-scenarios based on AWM: (C) S3<sub>Top1%</sub>, (D) S3<sub>Top2%</sub>, (E) S3<sub>Top5%</sub>, (F) S3<sub>Top10%</sub>, in which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively; (G) S4 (AWM-WssGBLUP method considering zero edges in the normalization).



Supplementary Figure 8: Manhattan plots of proportion of explained genetic variance by SNPs for skatole levels in a multi-line population in scenarios: (A) S1 (ssGBLUP); (B) S2 (WssGBLUP) and sub-scenarios based on AWM: (C) S3<sub>Top1%</sub>, (D) S3<sub>Top2%</sub>, (E) S3<sub>Top5%</sub>, (F) S3<sub>Top10%</sub>, in which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively; (G) S4 (AWM-WssGBLUP method considering zero edges in the normalization).



Supplementary Figure 9: Manhattan plots of proportion of explained genetic variance by SNPs for indole levels in a single-line population in scenarios: (A) S1 (ssGBLUP); (B) S2 (WssGBLUP) and sub-scenarios based on AWM: (C) S3<sub>Top1%</sub>, (D) S3<sub>Top2%</sub>, (E) S3<sub>Top5%</sub>, (F) S3<sub>Top10%</sub>, in which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively; (G) S4 (AWM-WssGBLUP method considering zero edges in the normalization).



Supplementary Figure 10: Manhattan plots of proportion of explained genetic variance by SNPs for indole levels in a multi-line population in scenarios: (A) S1 (ssGBLUP); (B) S2 (WssGBLUP) and sub-scenarios based on AWM: (C) S3<sub>Top1%</sub>, (D) S3<sub>Top2%</sub>, (E) S3<sub>Top5%</sub>, (F) S3<sub>Top10%</sub>, in which 1%, 2%, 5% and 10% top SNPs that explained the highest proportion of variance were used, respectively; (G) S4 (AWM-WssGBLUP method considering zero edges in the normalization).

## CHAPTER 4

### **Weighted genome-wide association study reveals new candidate genes related to steroid hormones potentially linked to boar taint**

#### **4.1. Abstract**

Androstenone (AND), skatole (SKA) and indole (IND) deposition in pig adipose tissue may cause the boar taint, an unpleasant taste and smell observed pork at cooking. The single-step genome-wide association study (WssGWAS) may help to understand the genetic mechanisms involved in the boar taint appearance. Therefore, we aimed to search genes potentially associated to boar taint appearance that were linked to SNPs identified by WssGWAS analyses; further, we aimed to investigate the biological processes relevant to boar taint appearance in which these gene are involved through gene network analyses. The WssGWAS was performed for AND, SKA and IND, using 4,922 pig phenotypes and 3,749 genotypes obtained using the Illumina PorcineSNP60 BeadChip. For each boar taint compound SNP windows that explained 0.5% or more of the total genetic variance were selected to search boar taint candidate genes. Then, the find genes were used to build the gene ontology network. Relevant QTL regions were found on SSC1 (SSC for *Sus scrofa* chromosome), 2, 3, 5, 6, 10, 12, 13, 15, and 17. We found six candidate genes, which are involved in different biological processes that may be related to boar taint appearance. Some of these gene ontologies identified were: response to gonadotropin-releasing hormone; steroidal hormone relate process; steroid hormone biosynthetic process; regulation of steroid metabolic process; regulation of steroid biosynthetic process; response to testosterone; intestinal absorption. These biological processes are relevant since skatole and indole produced and absorbed in the hind-gut, moreover, androstenone is a steroid hormone that, as other steroidal hormones, affects the skatole and indole metabolic degradation in the liver. Summarizing, this study identified the *HSD17B2* gene that was previously describe as linked to boar taint appearance and five new candidate genes with potential to explain boar taint phenotypes: *CRHBP*, *CTDSP2*, *CDK4*, *CYP27B1* e *SDR4E1*. These genes were mainly involved to biosynthesis, releasing and response to steroid hormones and intestinal absorption and may be possibly associated with boar taint compounds in the carcass.

Keywords: Androstenone. Candidate gene. Indol. Skatole. Steroid hormones



## 4.2. Introduction

The meat from adult non-castrated male pigs may present, at cooking, unpleasant taste and smell, which is known as boar taint. It happens because of lipophilic compounds deposited in adipose tissue, mainly skatole (3-methylindole), androstenone ( $5\alpha$ -androst-16-ene-3-one) and indole (4-phenyl-3-butenone, p-cresol and 4-ethylpheno). Androstenone is a steroid hormone produced and secreted by testis, while skatole and indole are produced by tryptophan bacterial degradation in hind-gut (Aldal et al., 2005; Aluwé et al., 2011; Babol et al., 2002; Claus et al., 1994).

All pigs present skatole and indole production in hind-gut therefore there is a strong genetic correlation between them (Grindflek et al., 2011; Lee et al., 2005). In non-castrated pigs, skatole and indole are not completely degraded by liver due to androstenone antagonism (Aluwé et al., 2011; Doran et al., 2002; Squires and Lundström, 1997; Zamaratskaia and Squires, 2009) which increase their deposition in boar adipose tissues. Probably different genetic mechanisms, that are not completely elucidated, are involved in the androstenone pathway and skatole and indole catabolism (Zamaratskaia and Squires, 2009).

Associations between boar taint compounds (androstenone, skatole and indole) and SNPs (single nucleotide polymorphisms) markers have been previously reported (Campos et al., 2015; Duijvesteijn et al., 2014, 2010). However, these previous studies identified different quantitative trait loci (QTL) regions and/or candidate genes (Drag et al., 2017, 2019; Duijvesteijn et al., 2014, 2010; Wang and Kadarmideen, 2019). Usually, genome-wide association studies (GWAS) of boar taint compounds have limited number of genotyped and phenotyped animals, which reduce its power. Single-step genomic best linear unbiased prediction (ssGBLUP) may be an alternative to increase the dataset used in association studies, because it allows the use of phenotypes of animals without genotypes. The ssGBLUP approach provides genomic estimated breeding values (GEBV) which can be used to estimate the SNP effects (Zhang et al., 2016). To improve the use of ssGBLUP for GWAS, Wang et al. (2012) proposed the SNPs weighting in the construction of the relationship matrix, according to its relevance for the analyzed trait, this methodology was called weighted single-step GWAS (WssGWAS). In the WssGWAS the SNP are weighted according to their explained genetic variance improving the resolution of GWAS to more precisely identify QTL (Lourenco et al., 2017).

Few studies analyze large datasets to identify novel QTL regions and to provide a deeper knowledge of the genes that control boar taint appearance. This may be due to the complexity of these traits, or the few phenotyped and genotyped animals or else to the great divergences between genetic groups (Lee et al., 2005; Moe et al., 2009; Wang and Kadarmideen, 2019). In this sense, the WssGWAS allows increase the boar taint data set and focus on the SNPs that actually are important to share for candidate genes.

Most of GWAS for boar taint end with the identification of the QTL regions and genes. Realizing the Gene ontologies (GO) study allows better understand of functional categories associated with annotated genes (Tang et al., 2007). The GO study from a set of identified genes help to understand GWAS results and elucidate the genetic control of the studied traits (Verardo et al., 2016, 2015). Therefore, we aimed to search genes potentially associated to boar taint appearance that were linked to SNPs identified by WssGWAS analyses; further, we aimed to investigate the biological processes relevant to boar taint appearance in which these gene are involved through gene network analyses.

#### **4.3. Materials and methods**

The data used for this study were obtained as part of routine data recording in a commercial breeding program. Samples collected for DNA extraction were only used for routine diagnostic purpose of the breeding program. Data recording and sample collection were conducted strictly in line with the rules given by Dutch Animal Research Authorities.

##### ***Phenotypic and genotypic data***

Data from three pig sire lines (L1: Duroc-based line; L2: synthetic line; L3: Pietrain) were used in the analyses, as described in Table 1. Boar taint compounds levels approximately followed log-normal distributions (Duijvesteijn et al., 2010; Mathur et al., 2014), therefore, the phenotypic information consisted of log-transformed levels of androstenone, skatole and indole (AND, SKA and IND, respectively) measured in the adipose tissue of 4,922 pig carcasses. The animals were slaughtered at approximately 177 ( $\pm$  9.9) days of age and boar taint compounds were measured in fat samples from the neck collected at the left carcass side, as described in Mathur et al. (2014). Briefly, androstenone concentration was determined using liquid chromatography-mass spectrometry (Verheyden et al., 2007), whereas indole and skatole contents were measured using fluorescence at 285 and 340 nm (Ampuero Kragten et al., 2011).

Table 1. Description of the number of animals with phenotypic and/or genotypic data from three sire lines.

| Line <sup>1</sup> | Animals    |           |                        |
|-------------------|------------|-----------|------------------------|
|                   | Phenotyped | Genotyped | Phenotyped + Genotyped |
| L1                | 3,572      | 1,316     | 854                    |
| L2                | 712        | 1,080     | 232                    |
| L3                | 638        | 1,353     | 123                    |
| Total             | 4,922      | 3,749     | 1,209                  |

<sup>1</sup>L1: Duroc-based line, L2: Synthetic line and L3: Pietrain.

Genotypic information of 3,749 animals from the three evaluated lines, genotyped using the Illumina PorcineSNP60 BeadChip, was also available for this study (Table 1). The genotypic data were submitted to quality control within line, in which we excluded SNPs located in both sex chromosomes, with call-rate smaller than 95%, MAF smaller than 1% and/or with strong deviations from the Hardy-Weinberg equilibrium ( $P < 10^{-7}$ ). A total of 43,375 SNPs were retained for further analysis after quality control. The pedigree included 13,604 animals.

### *Statistical analyses*

A weighted ssGWAS was performed using the BLUPF90 family of programs (Miszta et al., 2002) considering the same genetic and residual parameter as estimated in the chapter three. The analyses were conducted according to the following single trait mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

wherein:  $\mathbf{y}$  is a vector containing the logarithm of androstenone, skatole or indole levels;  $\mathbf{X}$  is the incidence matrix of fixed effects;  $\boldsymbol{\beta}$  is a vector of fixed effects, containing the effects of contemporary group (farm-year-month of slaughter), the covariates age at slaughter, scaled hot carcass weight at slaughter in each line and the line effect;  $\mathbf{Z}$  is the incidence matrix of animal additive genetic effect;  $\mathbf{a}$  is a vector of animal additive genetic effect,  $\mathbf{a} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{H})$ ;  $\mathbf{e}$  is a vector of residual effects,  $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$ ;  $\sigma_a^2$  and  $\sigma_e^2$  are the additive genetic and residual variances, respectively;  $\mathbf{I}$  is an identity matrix;  $\mathbf{H}$  is the relationship matrix based on both pedigree and genomic information, which inverse ( $\mathbf{H}^{-1}$ ) was given by Legarra et al. (2014):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

wherein  $\mathbf{A}^{-1}$  is the inverse of pedigree-based relationship matrix ( $\mathbf{A}$ );  $\mathbf{G}^{-1}$  is the inverse of the genomic relationship matrix;  $\mathbf{A}_{22}^{-1}$  is the inverse of pedigree-based relationship matrix from genotyped animals.

The  $\mathbf{G}$  matrix was calculated according to VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{ZDZ}'}{\sum_{i=1}^M 2p_i(1 - p_i)}$$

wherein  $\mathbf{Z}$  is a zero-centered matrix obtained by  $\mathbf{Z} = \mathbf{M} - \mathbf{P}$ , wherein  $\mathbf{M}$  is a  $m \times n$  (number of markers x number of animals) matrix, which specifies each individual genotype and  $\mathbf{P}$  is a matrix with the allele frequencies expressed as a difference of 0.5 and multiplied by 2, i.e. the  $i$  column of  $\mathbf{P}$  is given by  $2(p_i - 0.5)$ ;  $\mathbf{D}$  is a diagonal matrix, which will be better defined below;  $p_i$  and  $q_i$  are the SNP allelic frequencies in  $i^{\text{th}}$  loci.

The substitution effects of SNPs were computed using the weighted single-step GWAS proposed by Wang et al. (2012). In this method, the breeding values obtained through the ssGBLUP are used to calculate SNP effects, which in turn are applied in the computation of the variance explained by each marker. Then, these variances are used to build the D matrix iteratively. Three iterations were performed since it has been shown that it maximize genomic predictitive ability and correctly identify major SNPs (Lourenco et al., 2017; Zhang et al., 2016).

The percentage of genetic variance explained by the  $i$ -th set of consecutive SNPs ( $i$ -th SNP window) was calculated as described by Wang et al. (2014), using a window of 0.4 Mb, which is the average haplotype block size in commercial pig lines (Veroneze et al., 2014) including the lines considered in the present study;

For each boar taint compound (AND, SKA and IND), the SNP windows that explained 0.5% or more of the total genetic variance were selected for candidate genes search. The threshold of 0.5% was chosen based on the expected contribution of SNP windows (Marques et al., 2018; Sollero et al., 2017). In brief, assuming an equal contribution of the 3,930 windows in our data, the expected proportion of genetic variance explained by each window was 0.025%. Thus, we used a threshold of 0.5% which is equal to 20 times the expected variance.

We used the Gene database for *Sus scrofa* (Sscrofa11.1) available at National Center for Biotechnology Information (NCBI, 2019) to identify the genes located inside each selected window. Then, we used the Cytoscape software (Shannon et al., 2003) and ClueGO + CluePedia plug-in (Bindea et al., 2009) to build the gene ontology network. Briefly, ClueGO takes one or

more set of genes and search for related GO terms or Pathways based on hypergeometric test and Bonferroni correction. The software establish edges between genes and the chosen term based on a human Database. Thus, we were able to obtain gene networks highlighting biological roles potentially associated with boar taint.

#### 4.4.Results

A total of 16 windows explaining 0.5% or more of the genetic variance were identified (5, 4 and 7 for AND, SKA and IND, respectively). Together, these five windows for AND explained 2.90% of the genetic variance, the four windows identified for SKA explained 2.15% of the genetic variance and the seven windows identified for IND explained 4.47% of the genetic variance (Figure 1). We found only one overlapping QTL regions between SKA and IND in SSC10.

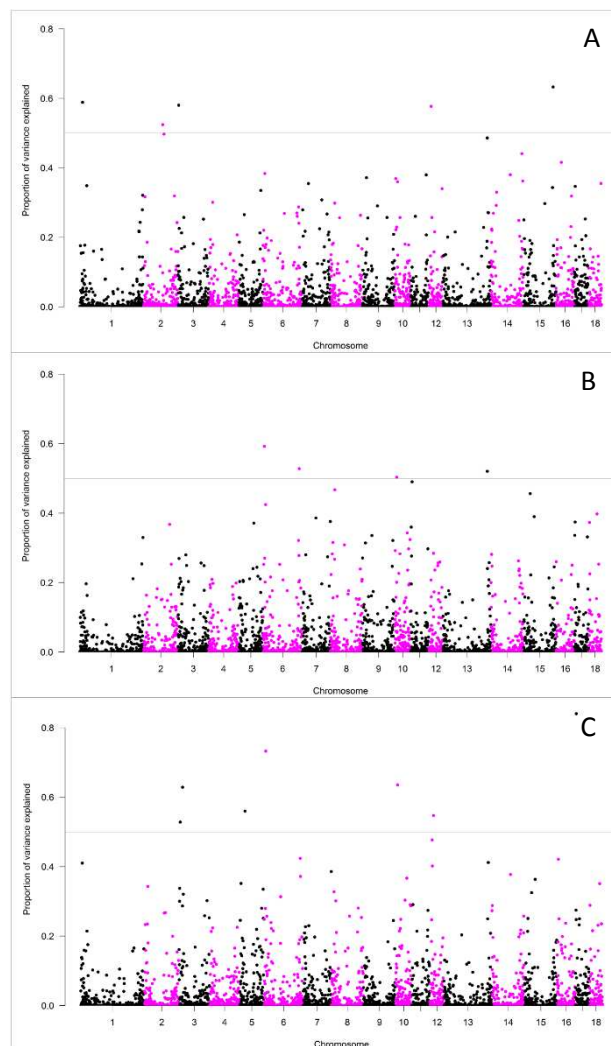


Figure 1 Proportion of explained genetic variance by each 0.4 MB windows for androstenone (A), skatole (B) and indole (C)

The relevant QTL regions were found on SSC1 (SSC for *Sus scrofa* chromosome), 2, 3, 5, 6, 10, 12, 13, 15, and 17 (Table 3). We found a total of 128 genes (Suppl. Table 1) located in these QTL regions, however, most of these genes have not been previously reported as directly associated with boar taint compounds. Using the gene network approach we were able to identify six potential candidate genes according to the biological process on which they are involved (Table 4).

Table 3 QTL regions identified for boar taint compounds

| Chr <sup>1</sup> | QTL region <sup>2</sup><br>(Mb) | Nb<br>SNP <sup>3</sup> | Var (%) <sup>4</sup> |                | Var (%) |                | Candidate gene <sup>6</sup> |
|------------------|---------------------------------|------------------------|----------------------|----------------|---------|----------------|-----------------------------|
|                  |                                 |                        | AND <sup>5</sup>     | SKA            | IND     |                |                             |
| 1                | 11.79 - 12.59                   | 18                     | 0.59                 | - <sup>7</sup> | -       | - <sup>8</sup> |                             |
| 2                | 85.05 - 85.85                   | 11                     | 0.52                 | -              | -       | -              | <i>CRHBP</i>                |
| 3                | 2.62 - 3.42                     | 23                     | 0.58                 | -              | -       | -              |                             |
| 3                | 5.04 - 5.84                     | 17                     | -                    | -              | 0.53    | -              |                             |
| 3                | 15.21 - 16.01                   | 15                     | -                    | -              | 0.63    | -              |                             |
| 5                | 22.34 - 23.14                   | 5                      | -                    | -              | 0.56    | -              | <i>CDK4/CYP27B1/CTDSP2</i>  |
| 6                | 5.94 - 6.74                     | 14                     | -                    | 0.59           | -       | -              | <i>HSD17B2/SDR42E1</i>      |
| 6                | 7.75 - 8.55                     | 24                     | -                    | -              | 0.73    | -              |                             |
| 6                | 157.08 - 157.88                 | 15                     | -                    | 0.53           | -       | -              |                             |
| 10               | 8.52 - 9.32                     | 13                     | -                    | -              | 0.64    | -              |                             |
| 10               | 8.91 - 9.71                     | 11                     | -                    | 0.50           | -       | -              |                             |
| 12               | 9.70 - 10.50                    | 11                     | 0.58                 | -              | -       | -              |                             |
| 12               | 15.68 - 16.48                   | 13                     | -                    | -              | 0.55    | -              |                             |
| 13               | 191.88 - 192.68                 | 18                     | -                    | 0.52           | -       | -              |                             |
| 15               | 126.99 - 127.79                 | 24                     | 0.63                 | -              | -       | -              |                             |
| 17               | 4.12 - 4.92                     | 11                     | -                    | -              | 0.84    | -              |                             |

<sup>1</sup> Chromosome

<sup>2</sup> Position of QTL region

<sup>3</sup> Number of SNPs within the windows

<sup>4</sup> Percentage of genetic variance explained by the windows

<sup>5</sup> Boar taint compounds: AND: androstenone; SKA: skatole; IND: indole

<sup>6</sup> Candidate gene(s) identified in the region

<sup>7</sup> The percentage of genetic variance explained by the QTL region is < 0.5%.

<sup>8</sup> No candidate genes associated with the trait.

Considering the QTLs regions associated to AND, we found the *CRHBP* gene involved with the response to gonadotropin-releasing hormone. From SKA results, the *HSD17B2* gene can be highlighted, since it is related to several sexual hormone related process. Finally, in the QTL regions for IND, genes related to hormone biosynthetic process, regulation of steroid

metabolic process and regulation of steroid biosynthetic process (*CTDSP2*, *CYP27B1*) were identified. In addition, the *CDK4* gene which is involved to response to testosterone and the *CYP27B1* involved to intestinal absorption were also identified.

Table 4 QTL regions identified for boar taint compounds

| <b>Genes</b>   | <b>Biological Process</b>                           | <b>Gene Ontology ID</b> |
|--|---|-------------------------|
| <i>CTDSP2</i> ; <i>CYP27B1</i> ; <i>HSD17B2</i> ; <i>SDR42E1</i> | steroid biosynthetic process                        | GO:0006694              |
| <i>CTDSP2</i>  | C21-steroid hormone biosynthetic process            | GO:0006700              |
| <i>CTDSP2</i>  | progesterone biosynthetic process                   | GO:0006701              |
| <i>HSD17B2</i>   | estrogen biosynthetic process                       | GO:0006703              |
| <i>CYP27B1</i>   | steroid catabolic process                           | GO:0006706              |
| <i>CTDSP2</i> ; <i>CYP27B1</i> ; <i>HSD17B2</i> ; <i>SDR42E1</i> | steroid metabolic process                           | GO:0008202              |
| <i>CTDSP2</i>  | C21-steroid hormone metabolic process               | GO:0008207              |
| <i>HSD17B2</i>   | estrogen metabolic process                          | GO:0008210              |
| <i>CYP27B1</i>   | negative regulation of steroid biosynthetic process | GO:0010894              |
| <i>CTDSP2</i> ; <i>CYP27B1</i>                                   | regulation of steroid metabolic process             | GO:0019218              |
| <i>CDK4</i>  | response to testosterone                            | GO:0033574              |
| <i>CTDSP2</i> ; <i>HSD17B2</i>                                   | cellular hormone metabolic process                  | GO:0034754              |
| <i>CTDSP2</i> ; <i>HSD17B2</i>                                   | hormone biosynthetic process                        | GO:0042446              |
| <i>CTDSP2</i>  | progesterone metabolic process                      | GO:0042448              |
| <i>CYP27B1</i>   | negative regulation of steroid metabolic process    | GO:0045939              |
| <i>CTDSP2</i>  | positive regulation of steroid metabolic process    | GO:0045940              |
| <i>CRHBP</i>   | negative regulation of hormone secretion            | GO:0046888              |
| <i>CYP27B1</i>   | regulation of steroid biosynthetic process          | GO:0050810              |
| <i>CYP27B1</i>   | intestinal absorption                               | GO:0050892              |
| <i>CRHBP</i>   | endocrine hormone secretion                         | GO:0060986              |
| <i>CRHBP</i>   | cellular response to estrogen stimulus              | GO:0071391              |
| <i>CRHBP</i>   | cellular response to estradiol stimulus             | GO:0071392              |
| <i>CDK4</i>  | cellular response to fatty acid                     | GO:0071398              |
| <i>CRHBP</i>   | response to gonadotropin-releasing hormone          | GO:0097210              |
| <i>CRHBP</i>   | cellular response to gonadotropin-releasing hormone | GO:0097211              |

The genes *HSD17B2*, *CTDSP2*, *CYP27B1* and *SDR4E1* are mainly related to steroid biosynthetic process (Figure 2), a very enriched biological processes potentially associated with the boar taint.

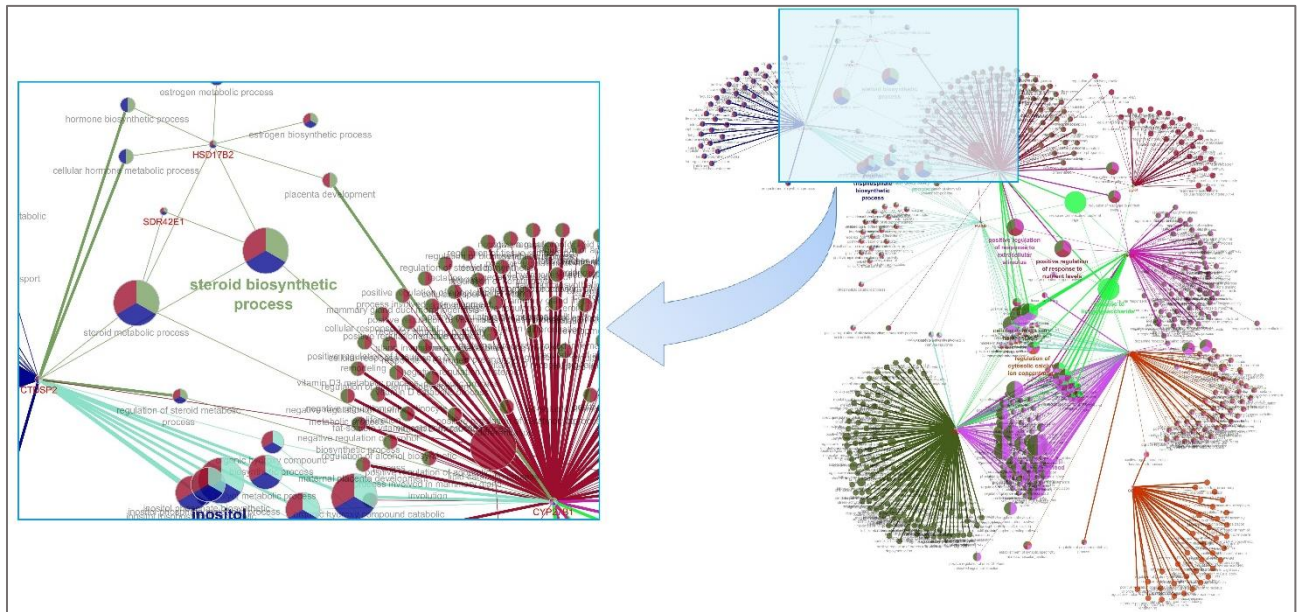


Figure 2 Gene network of biological process. Circle represent the genes and biological process. Edges represent the link between genes and biological process. Circle sizes are related to the enrichment of biological processes

#### 4.5. Discussion

Despite the high heritabilities for boar taint compounds estimated before (chapter three), the proportion of the additive genetic variance explained by each genome window was small, corroborating the highly polygenic profile of these traits (Lee et al., 2005; Quintanilla et al., 2003; Varona et al., 2005). Therefore, we found only 16 windows explaining more than 0.5% of genetic variance at least one boar taint compounds studied here. Previous studies reported some markers explaining higher proportion of variance for boar taint, however using only one genetic group and different approaches (Drag et al., 2018, 2017, 2019; Duijvesteijn et al., 2010; Wang and Kadarmideen, 2019).

The investigation of these 16 windows allowed the identification of 128 genes, which only one was previously reported in the literature as directly associated with boar taint compounds. Gene ontology networks may be helpful in the identification of gene candidates and for a better understand of the molecular mechanisms behind complex traits (Verardo et al., 2013) as boar taint. Our gene network analyses of biological processes allowed the identification of several candidate gene for boar taint compounds.

The *HSD17B2* (estradiol 17-beta-dehydrogenase) gene was previously reported in GWAS as associated with boar taint appearance (Moe et al., 2009; Rowe et al., 2014). This gene was linked to biological processes involving steroidal hormones biosynthesis in our gene network. The *HSD17B2* is located in a region on chromosome 6 that was previously found to



be significant for skatole level in Landrace pigs (Ramos et al., 2011) and is abundantly expressed in pig liver (Rowe et al., 2014). In addition, the *HSD17B2* gene in human, is involved with synthesis of 17 beta-hydroxysteroids (Labrie et al., 1995) that partially share same androstenone biosynthetic pathway in pigs (Chen et al., 2007).

We also identified the gene *CYP27B1* (cytochrome P450 family 27 subfamily B member 1) which belongs to the cytochrome P450 (CYP) family that acts directly in several biological process as metabolism and synthesis of cholesterol, steroids and other lipids. The CYP family also is involved in an oxidative phase of skatole degradation (Rowe et al., 2014; Zadinová et al., 2016). The *CYP27B1* were also associated to biological processes linked to steroidal hormone relate process and intestinal absorption in our gene network. These process may be involved to boar taint compounds since skatole and indole are exogenous molecules produced by tryptophan degradation in the hind-gut and absorbed by the intestine (Claus et al., 1994; Laderoute et al., 2019; Rius et al., 2005; Zamaratskaia et al., 2004; Zamaratskaia and Squires, 2009). Moreover, androstenone is an steroid hormone that, as other steroidal hormones, affects the skatole and indole metabolic clearance by the liver (Zamaratskaia and Squires, 2009). Thus, the *CYP27B1* is a strong candidate gene for boar taint compounds deposition.

Our gene network reveled four more potentially genes candidates to boar taint compounds. One of these genes is *CRHBP* (Corticotropin Releasing Hormone Binding Protein), which encodes a protein that inactivates the Corticotropin-releasing hormone and may prevent inappropriate pituitary-adrenal stimulation (Perkins et al., 1995). This gene may indirectly act on AND levels, since adrenal may convert pregnenolone into androst-16-ene steroids (Meadus et al., 1993) and changes gonadotropin release by pituitary gland (e.g. immunocastration) which affect testicular function in males (Ayalew, 2019) and consequently affects the AND and SKA levels in pig fat (Poulsen Nautrup et al., 2018)

The genes *CYP27B1* and *CTDSP2* (C-terminal domain small phosphatase 2) were related to Regulation of Steroid Biosynthetic Processes and Regulation of Steroid Metabolic Processes. The *CTDSP2* gene are C-class phosphatases expressed in diverse organs and induces the neuronal differentiation (Han et al., 2012). *CTDSP2* activity was described as promoter clearance during steroid-activated transcription (Yilmaz et al., 2010) which may justify its association to boar taint.

Another gene linked to steroid hormones identified was *CDK4* (Cyclin Dependent Kinase 4) gene. This gene expression is related to Leydig cells development, being an important member of the Ser/Thr protein kinase family indispensable for cell cycle G1 phase progression

(DONG et al., 2007) affecting Sertoli cells activity and spermatogenesis in monkeys (Xin-Chang et al., 2002). Therefore, it is plausible that this gene may be linked to the IND level since the AND produced by the Leydig cells together with testicular steroids (Laderoute et al., 2019) has antagonistic action to the IND degradation.

The most enriched biological process linked to boar taint identified was steroid biosynthetic process which was sheared by *HSD17B2*, *CTDSP2*, *CYP27B1* and *SDR4E1* (also named *TSTA3* - tissue specific transplantation antigen P35B). Although IND and SKA are not steroid hormones, the association with this biological process is compatible since their deposition in adipose tissue are strongly influenced by the androstenone (an steroidal hormone) level (Aluwé et al., 2011; Doran et al., 2002; Zamaratskaia and Squires, 2009). Moreover, several studies have been demonstrated that genetic correlations between androstenone and skatole or indole are moderate, around 0.31 and 0.46 (Campos et al., 2015; Lee et al., 2005; Windig et al., 2012) evidencing that some genes may affect these traits.

#### **4.6. Conclusion**

In summary, this study identified the *HSD17B2* gene that was previously described as linked to boar taint appearance. New candidate genes with potential to explain boar taint phenotypes were found: *CRHBP*, *CTDSP2*, *CDK4*, *CYP27B1* e *SDR4E1*. These genes were mainly involved to biosynthesis, releasing and response to steroid hormones and intestinal absorption and may be possibly associated with boar taint compounds in the carcass. More post-GWAS studies may be helpful to better understand the genetic mechanisms in which these genes are involved with boar taint appearance.

#### **4.7. References**

- Aldal, I., Ø. Andresen, A.K. Egeli, J. Haugen, A. Grødum, et al. 2005. Levels of androstenone and skatole and the occurrence of boar taint in fat from young boars. *Livest Prod Sci* 95(1–2): 121–129. doi: 10.1016/j.livprodsci.2004.12.010.
- Aluwé, M., S. Millet, K.M. Bekaert, F.A.M.M. Tuytens, L. Vanhaecke, et al. 2011. Influence of breed and slaughter weight on boar taint prevalence in entire male pigs. *Animal* 5(8): 1283–1289. doi: 10.1017/S1751731111000164.
- Ampuero Kragten, S., B. Verkuylen, H. Dahlmans, M. Hortos, J.A. Garcia-Regueiro, et al. 2011. Inter-laboratory comparison of methods to measure androstenone in pork fat. *Animal* 5(10): 1634–1642. doi: 10.1017/S1751731111000553.
- Ayalew, G. 2019. A Review on the Effect of Immunocastration Against Gonadal Physiology and Boar Taint (Intergovernmental Panel on Climate Change, editor). *Biomed Nurs* 5(1): 1–30. doi: 10.7537/marsbnj050119.03.

- Babol, J., E.J. Squires, and E.A. Gullett. 2002. Factors affecting the level of boar taint in entire male pigs as assessed by consumer sensory panel. *Meat Sci* 61: 33–40. doi: 10.1016/s0309-1740(01)00159-0.
- Bindea, G., B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, et al. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25(8): 1091–1093. doi: 10.1093/bioinformatics/btp101.
- Campos, C.F. de, M.S. Lopes, F.F. e Silva, R. Veroneze, E.F. Knol, et al. 2015. Genomic selection for boar taint compounds and carcass traits in a commercial pig population. *Livest Sci* 174: 10–17. doi: 10.1016/j.livsci.2015.01.018.
- Chen, G., E. Bourneuf, S. Marklund, G. Zamaratskaia, A. Madej, et al. 2007. Gene expression of 3 $\beta$ -hydroxysteroid dehydrogenase and 17 $\beta$ -hydroxysteroid dehydrogenase in relation to androstenone, testosterone, and estrone sulphate in gonadally intact male and castrated pigs. *J Anim Sci* 85(10): 2457–2463. doi: 10.2527/jas.2007-0087.
- Claus, R., U. Weiler, and A. Herzog. 1994. Physiological Aspects of Androstenone and Skatole Formation in the Boar A Review with Experimental Data. *Meat Sci* 38: 289–305. doi: 10.1016/0309-1740(94)90118-X.
- DONG, L., S.A. JELINSKY, J.N. FINGER, D.S. JOHNSTON, G.S. KOPF, et al. 2007. Gene Expression During Development of Fetal and Adult Leydig Cells. *Ann N Y Acad Sci* 1120(1): 16–35. doi: 10.1196/annals.1411.016.
- Doran, E., F.W. Whittington, J.D. Wood, and J.D. Mcgivan. 2002. Cytochrome P450IIE1 ( CYP2E1 ) is induced by skatole and this induction is blocked by androstenone in isolated pig hepatocytes. *Chem Biol Interact* 140: 81–92. doi: 10.1016/S0009-2797(02)00015-7.
- Drag, M., M.B. Hansen, and H.N. Kadarmideen. 2018. Systems genomics study reveals expression quantitative trait loci, regulator genes and pathways associated with boar taint in pigs. *PLoS One* 13(2): 1–30. doi: 10.1371/journal.pone.0192673.
- Drag, M.H., L.J.A. Kogelman, H. Maribo, L. Meinert, P.D. Thomsen, et al. 2019. Characterization of eQTLs associated with androstenone by RNA sequencing in porcine testis. *Physiol Genomics* 51(10): 488–499. doi: 10.1152/physiolgenomics.00125.2018.
- Drag, M., R. Skinkyté-Juskiené, D.N. Do, L.J.A. Kogelman, and H.N. Kadarmideen. 2017. Differential expression and co-expression gene networks reveal candidate biomarkers of boar taint in non-castrated pigs. *Sci Rep* 7(1): 1–18. doi: 10.1038/s41598-017-11928-0.
- Duijvesteijn, N., E.F. Knol, and P. Bijma. 2014. Boar taint in entire male pigs : A genomewide association study for direct and indirect genetic effects on androstenone. *J Anim Sci* 92: 4319–4328. doi: 10.2527/jas2014-7863.
- Duijvesteijn, N., E.F. Knol, J.W.M. Merks, R.P.M.A. Crooijmans, M.A.M. Groenen, et al. 2010. A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. *BMC Genet* 11(42): 1–11. doi: 10.1186/1471-2156-11-42.
- Grindflek, E., T.H.E. Meuwissen, T. Aasmundstad, H. Hamland, M.H.S. Hansen, et al. 2011. Revealing genetic relationships between compounds affecting boar taint and reproduction in pigs. *J Anim Sci* 89(3): 680–692. doi: 10.2527/jas.2010-3290.
- Han, J., A.M. Denli, and F.H. Gage. 2012. The enemy within: intronic miR-26b represses its host gene, *ctdsp2*, to regulate neurogenesis. *Genes Dev* 26(1): 6–10. doi:

10.1101/gad.184416.111.

- Labrie, Y., F. Durocher, Y. Lachance, C. Turgeon, J. Simard, et al. 1995. The Human Type II 17 $\beta$ -Hydroxysteroid Dehydrogenase Gene Encodes Two Alternatively Spliced mRNA Species. *DNA Cell Biol* 14(10): 849–861. doi: 10.1089/dna.1995.14.849.
- Laderoute, H., C. Bone, D. Brewer, and E.J. Squires. 2019. The synthesis of 16-androstene sulfoconjugates from primary porcine Leydig cell culture. *Steroids* 146(February): 14–20. doi: 10.1016/j.steroids.2019.03.007.
- Lee, G.J., A.L. Archibald, A.S. Law, S. Lloyd, J. Wood, et al. 2005. Detection of quantitative trait loci for androstenone, skatole and boar taint in a cross between Large White and Meishan pigs. *Anim Genet* 36(1): 14–22. doi: 10.1111/j.1365-2052.2004.01214.x.
- Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. *Livest Sci* 166(1): 54–65. doi: 10.1016/j.livsci.2014.04.029.
- Lourenco, D.A.L., B.O. Fragomeni, H.L. Bradford, I.R. Menezes, J.B.S. Ferraz, et al. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. *J Anim Breed Genet* 134(6): 463–471. doi: 10.1111/jbg.12288.
- Marques, D.B.D., J.W.M. Bastiaansen, M.L.W.J. Broekhuijse, M.S. Lopes, E.F. Knol, et al. 2018. Weighted single-step GWAS and gene network analysis reveal new candidate genes for semen traits in pigs. *Genet Sel Evol* 50(1): 1–14. doi: 10.1186/s12711-018-0412-z.
- Mathur, P.K., J. ten Napel, R.E. Crump, H.A. Mulder, and E.F. Knol. 2014. Genetic relationship between boar taint compounds, human nose scores, and reproduction traits in pigs. *J Anim Breed Genet* 91(9): 4080–4089. doi: <https://doi.org/10.2527/jas.2013-6478>.
- Meadus, W.J., J.I. Mason, and E.J. Squires. 1993. Cytochrome P450c17 from porcine and bovine adrenal catalyses the formation of 5,16-androstadien-3 $\beta$ -ol from pregnenolone in the presence of cytochrome b5. *J Steroid Biochem Mol Biol* 46(5): 565–572. doi: 10.1016/0960-0760(93)90183-W.
- Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, et al. 2002. BLUPF90 and related programs (BGF90). In *Proceedings of the 7th world congress on genetics applied to livestock production. Proceedings of the 7th world congress on genetics applied to livestock production. Montpellier-France. p. 19–23 Aug 2002*
- Moe, M., S. Lien, T. Aasmundstad, T.H. Meuwissen, M.H. Hansen, et al. 2009. Association between SNPs within candidate genes and compounds related to boar taint and reproduction. *BMC Genet* 10(1): 32. doi: 10.1186/1471-2156-10-32.
- NCBI. 2019. National Center for Biotechnology Information.
- Perkins, A. V, C.D.A. Wolfe, F. Eben, P. Soothill, and E.A. Linton. 1995. Corticotrophin-releasing hormone-binding protein in human fetal plasma. *J Endocrinol* 146(3): 395–401. doi: 10.1677/joe.0.1460395.
- Poulsen Nautrup, B., I. Van Vlaenderen, A. Aldaz, and C.K. Mah. 2018. The effect of immunization against gonadotropin-releasing factor on growth performance, carcass characteristics and boar taint relevant to pig producers and the pork packing industry: A meta-analysis. *Res Vet Sci* 119(June): 182–195. doi: 10.1016/j.rvsc.2018.06.002.
- Quintanilla, R., O. Demeure, J.P. Bidanel, D. Milan, N. Iannuccelli, et al. 2003. Detection of

- quantitative trait loci for fat androstenone levels in pigs. *J Anim Sci* 81(2): 385–394. doi: 10.2527/2003.812385x.
- Ramos, A.M., N. Duijvesteijn, E.F. Knol, J.W.M. Merks, H. Bovenhuis, et al. 2011. The distal end of porcine chromosome 6p is involved in the regulation of skatole levels in boars. *BMC Genet* 12. doi: 10.1186/1471-2156-12-35.
- Rius, M.A., M. Hortós, and J.A. García-Regueiro. 2005. Influence of volatile compounds on the development of off-flavours in pig back fat samples classified with boar taint by a test panel. *Meat Sci* 71(4): 595–602. doi: 10.1016/j.meatsci.2005.03.014.
- Rowe, S.J., B. Karacaören, D.-J. de Koning, B. Lukic, N. Hastings-Clark, et al. 2014. Analysis of the genetics of boar taint reveals both single SNPs and regional effects. *BMC Genomics* 15(1): 424. doi: 10.1186/1471-2164-15-424.
- Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, et al. 2003. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*: 2498–2504. doi: 10.1101/gr.1239303.metabolite.
- Sollero, B.P., V.S. Junqueira, C.C.G. Gomes, A.R. Caetano, and F.F. Cardoso. 2017. Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. *Genet Sel Evol* 49(1): 49. doi: 10.1186/s12711-017-0325-2.
- Squires, E.J., and K. Lundström. 1997. Relationship between cytochrome P450IIE1 in liver and levels of skatole and its metabolites in intact male pigs. *J Anim Sci* 75(9): 2506. doi: 10.2527/1997.7592506x.
- Tang, Z., Y. Li, P. Wan, X. Li, S. Zhao, et al. 2007. LongSAGE analysis of skeletal muscle at three prenatal stages in Tongcheng and Landrace pigs. *Genome Biol* 8(6): R115. doi: 10.1186/gb-2007-8-6-r115.
- VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. *J Dairy Sci* 91(11): 4414–4423. doi: 10.3168/jds.2007-0980.
- Varona, L., O. Vidal, R. Quintanilla, M. Gil, A. Sánchez, et al. 2005. Bayesian analysis of quantitative trait loci for boar taint in a Landrace outbred population. *J Anim Sci* 83(2): 301–307. doi: 10.2527/2005.832301x.
- Verardo, L.L., M.S. Lopes, S. Wijga, O. Madsen, F.F. Silva, et al. 2016. After genome-wide association studies: Gene networks elucidating candidate genes divergences for number of teats across two pig populations I. *J Anim Sci* 94(4): 1446–1458. doi: 10.2527/jas.2015-9917.
- Verardo, L.L., C.S. Nascimento, F.F. Silva, E. Gasparino, M.F. Martins, et al. 2013. Identification and validation of differentially expressed genes from pig skeletal muscle. *J Anim Breed Genet* 130(5): 372–381. doi: 10.1111/jbg.12006.
- Verardo, L.L., F.F. Silva, L. Varona, M.D. V. Resende, J.W.M. Bastiaansen, et al. 2015. Bayesian GWAS and network analysis revealed new candidate genes for number of teats in pigs. *J Appl Genet* 56(1): 123–132. doi: 10.1007/s13353-014-0240-y.
- Verheyden, K., H. Noppe, M. Aluwé, S. Millet, J. Vanden Bussche, et al. 2007. Development and validation of a method for simultaneous analysis of the boar taint compounds indole, skatole and androstenone in pig fat using liquid chromatography–multiple mass spectrometry. *J Chromatogr A* 1174(1–2): 132–137. doi: 10.1016/j.chroma.2007.08.075.
- Veroneze, R., J.W.M. Bastiaansen, E.F. Knol, S.E.F. Guimarães, F.F. Silva, et al. 2014.

- Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (*Sus scrofa*) populations. *BMC Genet* 15(1): 126. doi: 10.1186/s12863-014-0126-3.
- Wang, X., and H.N. Kadarmideen. 2019. Genome-wide DNA methylation analysis using next-generation sequencing to reveal candidate genes responsible for boar taint in pigs. *Anim Genet*. doi: 10.1111/age.12842.
- Wang, H., I. Misztal, I. Aguilar, A. Legarra, R.L. Fernando, et al. 2014. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Front Genet* 5(MAY): 1–10. doi: 10.3389/fgene.2014.00134.
- Windig, J.J., H.A. Mulder, J. ten Napel, E.F. Knol, P.K. Mathur, et al. 2012. Genetic parameters for androstenone, skatole, indole, and human nose scores as measures of boar taint and their relationship with finishing traits. *J Anim Sci* 90(7): 2120–2129. doi: 10.2527/jas.2011-4700.
- Xin-Chang, Z., W. Peng, H. Zhao-Yuan, H. Xiao-Bin, Z. Ru-Jin, et al. 2002. Expression of P16INK4a in testis of rhesus monkey during heat stress and testosterone undecanoate induced azoospermia or oligozoospermia☆. *Contraception* 65(3): 251–255. doi: 10.1016/S0010-7824(01)00305-5.
- Yilmaz, S., M. Boffito, S. Collot-Teixeira, F. De Lorenzo, L. Waters, et al. 2010. Investigation of low-dose ritonavir on human peripheral blood mononuclear cells using gene expression whole genome microarrays. *Genomics* 96(1): 57–65. doi: 10.1016/j.ygeno.2010.03.011.
- Zadinová, K., R. Stupka, A. Stratil, J. Čítek, K. Vehovský, et al. 2016. Boar taint – the effects of selected candidate genes associated with androstenone and skatole levels – a review. *Anim Sci Pap Reports* 34(2): 107–128.
- Zamaratskaia, G., J. Babol, H. Andersson, and K. Lundström. 2004. Plasma skatole and androstenone levels in entire male pigs and relationship between boar taint

#### 4.8. Supplementary material

Supplementary Table 1. Genes identified under the studied QTLs

| Chromosome   | Position of QTL region (Mb) | Identified gene | Gene start position (Mb) | Gene end position (Mb) |
|--------------|-----------------------------|-----------------|--------------------------|------------------------|
| 1            | 11.79 - 12.59               | TIAM2           | 11.65                    | 11.90                  |
|              |                             | SCAF8           | 11.89                    | 12.12                  |
|              |                             | CNKSR3          | 12.30                    | 12.40                  |
|              |                             | IPCEF1          | 12.51                    | 12.60                  |
|              |                             | LOC110259934    | 12.44                    | 12.50                  |
|              |                             | LOC110256595    | 12.12                    | 12.13                  |
| 2            | 85.05 - 85.85               | CRHBP           | 85.81                    | 85.83                  |
|              |                             | F2RL1           | 85.73                    | 85.75                  |
|              |                             | F2R             | 85.64                    | 85.66                  |
|              |                             | S100Z           | 85.76                    | 85.77                  |
|              |                             | F2RL2           | 85.53                    | 85.54                  |
|              |                             | LOC106509481    | 84.99                    | 85.00                  |
|              |                             | LOC110259403    | 85.27                    | 85.29                  |
|              |                             | LOC110259402    | 85.49                    | 85.51                  |
| 3            | 2.62 - 3.42                 | SDK1            | 2.81                     | 3.33                   |
|              |                             | LOC102158292    | 3.38                     | 3.48                   |
|              |                             | LOC110259838    | 3.33                     | 3.36                   |
| 3            | 5.04 - 5.84                 | NPTX2           | 5.72                     | 5.73                   |
|              |                             | BAIAP2L1        | 5.47                     | 5.57                   |
|              |                             | TECPR1          | 5.41                     | 5.44                   |
|              |                             | BRI3            | 5.46                     | 5.47                   |
|              |                             | LMTK2           | 5.32                     | 5.41                   |
|              |                             | CCZ1            | 5.20                     | 5.23                   |
|              |                             | AIMP2           | 5.11                     | 5.12                   |
|              |                             | ANKRD61         | 5.10                     | 5.10                   |
|              |                             | USP42           | 5.00                     | 5.05                   |
|              |                             | EIF2AK1         | 5.08                     | 5.11                   |
|              |                             | PMS2            | 5.12                     | 5.15                   |
|              |                             | BHLHA15         | 5.41                     | 5.41                   |
|              |                             | LOC100516852    | 5.15                     | 5.20                   |
|              |                             | LOC102163639    | 5.23                     | 5.23                   |
|              |                             | LOC106509627    | 5.27                     | 5.28                   |
|              |                             | LOC110259846    | 5.43                     | 5.44                   |
|              |                             | LOC110259847    | 5.58                     | 5.59                   |
| LOC106509632 | 5.61                        | 5.62            |                          |                        |
| LOC106509631 | 5.60                        | 5.61            |                          |                        |

|   |               |           |       |       |
|---|---------------|-----------|-------|-------|
|   |               | GALNT17   | 15.03 | 15.45 |
| 3 | 15.21 - 16.01 | CALN1     | 15.48 | 15.95 |
|   |               | TYW1      | 15.97 | 16.16 |
|   |               | MYO1A     | 22.34 | 22.36 |
|   |               | NAB2      | 22.40 | 22.41 |
|   |               | NEMP1     | 22.37 | 22.39 |
|   |               | LRP1      | 22.44 | 22.52 |
|   |               | NXPH4     | 22.52 | 22.54 |
|   |               | NDUFA4L2  | 22.54 | 22.55 |
|   |               | STAC3     | 22.55 | 22.56 |
|   |               | SHMT2     | 22.54 | 22.54 |
|   |               | GLI1      | 22.74 | 22.75 |
|   |               | DCTN2     | 22.80 | 22.81 |
|   |               | KIF5A     | 22.81 | 22.85 |
|   |               | DDIT3     | 22.79 | 22.79 |
|   |               | MARS1     | 22.75 | 22.79 |
|   |               | PIP4K2C   | 22.85 | 22.87 |
|   |               | ARHGAP9   | 22.75 | 22.76 |
| 5 | 22.34 - 23.14 | INHBC     | 22.72 | 22.73 |
|   |               | ARHGEF25  | 22.87 | 22.88 |
|   |               | INHBE     | 22.73 | 22.73 |
|   |               | DTX3      | 22.87 | 22.87 |
|   |               | SLC26A10  | 22.88 | 22.89 |
|   |               | MBD6      | 22.79 | 22.80 |
|   |               | B4GALNT1  | 22.89 | 22.90 |
|   |               | OS9       | 22.96 | 23.00 |
|   |               | CYP27B1   | 23.05 | 23.06 |
|   |               | EEF1AKMT3 | 23.06 | 23.09 |
|   |               | CTDSP2    | 23.11 | 23.13 |
|   |               | MIR26A    | 23.11 | 23.11 |
|   |               | MARCH9    | 23.05 | 23.05 |
|   |               | METTL1    | 23.06 | 23.06 |
|   |               | CDK4      | 23.04 | 23.04 |
|   |               | ATP23     | 23.18 | 23.28 |



|    |                 |              |       |       |
|----|-----------------|--------------|-------|-------|
|    |                 | HSD17B2      | 6.30  | 6.36  |
|    |                 | CMIP         | 6.65  | 6.88  |
|    |                 | MPHOSPH6     | 6.25  | 6.27  |
| 6  | 5.94 - 6.74     | SDR42E1      | 6.39  | 6.40  |
|    |                 | PLCG2        | 6.43  | 6.60  |
|    |                 | LOC102166616 | 5.86  | 6.25  |
|    |                 | LOC110260906 | 6.36  | 6.38  |
|    |                 | LOC106510465 | 7.92  | 8.34  |
| 6  | 7.75 - 8.55     | LOC106510470 | 8.52  | 8.53  |
|    |                 | LOC106510467 | 8.47  | 8.47  |
|    |                 | LOC106510468 | 8.38  | 8.39  |
| 6  | 157.08 - 157.88 | -            | -     | -     |
|    |                 | LOC106504446 | 8.62  | 8.63  |
|    |                 | LOC110255626 | 8.52  | 8.52  |
|    |                 | LOC102164422 | 8.54  | 8.95  |
| 10 | 8.52 - 9.32     | LYPLAL1      | 8.95  | 9.06  |
|    |                 | LOC106507893 | 9.21  | 9.22  |
|    |                 | LOC110255627 | 9.21  | 9.22  |
|    |                 | LOC110255735 | 9.11  | 9.11  |
|    |                 | LYPLAL1      | 8.95  | 9.06  |
|    |                 | LOC102164422 | 8.54  | 8.95  |
|    |                 | LOC110255735 | 9.11  | 9.11  |
|    |                 | LOC106507893 | 9.21  | 9.22  |
| 10 | 8.91 - 9.71     | LOC110255627 | 9.20  | 9.25  |
|    |                 | EPRS         | 9.58  | 9.65  |
|    |                 | LOC102165479 | 9.66  | 9.68  |
|    |                 | IARS2        | 9.69  | 9.74  |
|    |                 | SLC30A10     | 9.53  | 9.55  |
|    |                 | MIR215       | 9.70  | 9.70  |
|    |                 | KCNJ16       | 10.39 | 10.54 |
| 12 | 9.70 - 10.50    | KCNJ2        | 10.35 | 10.38 |
|    |                 | LOC102166398 | 10.45 | 10.46 |
|    |                 | LOC110255979 | 10.36 | 10.36 |

|    |                 |              |        |        |
|----|-----------------|--------------|--------|--------|
|    |                 | TANC2        | 15.81  | 15.45  |
|    |                 | MARCH10      | 16.05  | 15.97  |
|    |                 | TLK2         | 16.26  | 16.13  |
|    |                 | EFCAB3       | 16.37  | 16.31  |
|    |                 | MRC2         | 16.12  | 16.06  |
| 12 | 15.68 - 16.48   | LOC110256166 | 15.95  | 15.92  |
|    |                 | LOC102158113 | 15.89  | 15.89  |
|    |                 | LOC110256164 | 16.05  | 15.99  |
|    |                 | LOC100737967 | 16.24  | 16.23  |
|    |                 | LOC110256160 | 16.26  | 16.26  |
|    |                 | LOC102160182 | 16.45  | 16.45  |
|    |                 | LOC100516640 | 16.44  | 16.39  |
|    |                 | LOC106505844 | 192.07 | 191.96 |
|    |                 | N6AMT1       | 192.27 | 192.25 |
|    |                 | LTN1         | 192.36 | 192.30 |
| 13 | 191.88 - 192.68 | USP16        | 192.41 | 192.38 |
|    |                 | MAP3K7CL     | 192.53 | 192.49 |
|    |                 | CCT8         | 192.42 | 192.41 |
|    |                 | RWDD2B       | 192.37 | 192.36 |
|    |                 | BACH1        | 192.70 | 192.66 |
|    |                 | NYAP2        | 127.30 | 127.03 |
| 15 | 126.99 - 127.79 | LOC100512318 | 127.43 | 127.43 |
|    |                 | LOC100511416 | 127.44 | 127.44 |
|    |                 | LOC110256922 | 127.56 | 127.54 |
|    |                 | ZDHHC2       | 4.93   | 4.86   |
| 17 | 4.12 - 4.92     | MICU3        | 4.83   | 4.72   |
|    |                 | FGF20        | 4.70   | 4.69   |
|    |                 | LOC102161130 | 4.76   | 4.76   |

## GENERAL CONCLUSION

In summary, we verified that there is causal relationship regarding boar taint compounds in carcasses and biopsies. In general, the causal structure showed that skatole in the carcass is directly affected by androstenone and indole measured in biopsies and indirectly affected skatole measured in biopsies and androstenone in the carcass. Moreover, only skatole in carcass affects directly indole in the carcass.

The genomic prediction for boar taint compounds in carcass using the traditional single step GBLUP obtained slightly worse in predictive ability and bias than weighted methods. However, the traditional single step GBLUP resulted in the best predictive abilities and biases for androstenone. Despite a small improvement in the prediction was observed, the weighted prediction strategies increased the number of analyses steps, which may not justify their application.

Genes associated with steroid metabolism biosynthesis and intestinal absorption may be possibly associated with boar taint compounds in the carcass. These genes include the HSD17B2, previously reported as a candidate gene to boar taint, and five new candidate genes (CDK4, CRHBP, CTDSP2, CYP27B1 and SDR4E1) involved in biological process that may be related to boar taint compounds.