

**FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS  
UNIVERSIDADE ESTADUAL PAULISTA  
CÂMPUS DE JABOTICABAL**

**POWER OF QTL MAPPING OF DIFFERENT GENOME-WIDE  
ASSOCIATION METHODS FOR TRAITS UNDER  
DIFFERENT GENETIC STRUCTURES: A SIMULATION  
STUDY**

**Baltasar Fernandes Garcia Neto  
Zootecnista**

**2018**

**FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS  
UNIVERSIDADE ESTADUAL PAULISTA  
CÂMPUS DE JABOTICABAL**

**POWER OF QTL MAPPING OF DIFFERENT GENOME-WIDE  
ASSOCIATION METHODS FOR TRAITS UNDER  
DIFFERENT GENETIC STRUCTURES: A SIMULATION  
STUDY**

**Baltasar Fernandes Garcia Neto  
Advisor: Dr. Roberto Carneiro**

**Dissertation presented to The Faculdade de  
Ciências Agrárias e Veterinárias Unesp,  
Câmpus de Jaboticabal, in partial fulfillment  
of requirements for the degree of Mestre em  
Genética e Melhoramento Animal) (*Master in  
Animal Breeding and Genetics*)**

**2018**

Garcia Neto, Baltasar Fernandes  
G216p Power of qtl mapping of different genome-wide association  
methods for traits under different genetic structures: a simulation study  
/ Baltasar Fernandes Garcia Neto. -- Jaboticabal, 2018  
v, 22 p. : il. ; 28 cm

Dissertação (mestrado) - Universidade Estadual Paulista,  
Faculdade de Ciências Agrárias e Veterinárias, 2018  
Orientador: Roberto Carvalheiro  
Banca examinadora: Danisio Prado Munari, Ricardo Vieira Ventura  
Bibliografia

1. GWAS. 2. Genetic Structure. 3. wssGBLUP. I. Título. II.  
Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 636.082

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação –  
Diretoria Técnica de Biblioteca e Documentação - UNESP, Câmpus de Jaboticabal.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: POWER OF QTL MAPPING OF DIFFERENT GENOME-WIDE ASSOCIATION METHODS FOR TRAITS UNDER DIFFERENT GENETIC STRUCTURES: A SIMULATION STUDY

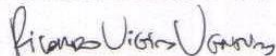
AUTOR: BALTASAR FERNANDES GARCIA NETO

ORIENTADOR: ROBERTO CARVALHEIRO

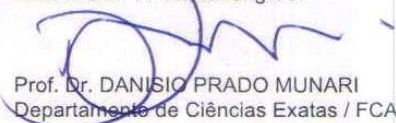
Aprovado como parte das exigências para obtenção do Título de Mestre em GENÉTICA E MELHORAMENTO ANIMAL, pela Comissão Examinadora:



Prof. Dr. ROBERTO CARVALHEIRO (Participação por Videoconferência)  
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal



Prof. Dr. RICARDO VIEIRA VENTURA  
FZEA / USP / Pirassununga/SP



Prof. Dr. DANISIO PRADO MUNARI  
Departamento de Ciências Exatas / FCAV / UNESP - Jaboticabal

Jaboticabal, 27 de fevereiro de 2018

## **DADOS CURRICULARES DO AUTOR**

Baltasar was born at Jaboticabal, Sao Paulo in October 7<sup>th</sup>, 1991. In 2010, he started the Animal Science undergraduate course at Universidade Estadual Paulista – UNESP in the same city. Since the beginning of the course, Baltasar was engaged in Aquaculture area, more specifically in shrimp and prawn farming. He participated in different post-graduate studies, and had a scientific research scholarship with Prof. Dr. Wagner Valenti, granted by the National Council for Scientific and Technological Development (CNPq). In 2013, he was granted with a scholarship (Science without Borders) provided by Coordination for the Improvement of Higher Education Personnel (CAPES) to study for one year in Ireland, at the National University of Galway – NUIG, where he was able to improve his English skills, attend some courses and learn more about a different culture. In 2015, he took his undergraduate degree in Animal Science at UNESP, defending a literature review research on heritability estimates in shrimp, under the supervision of Dr. Roberto Carneiro. In the same year, he spent 3 months at La Paz, Baja California Sur, Mexico doing his obligatory internship at CIBNOR research center, under the supervision of Prof. Dr. Ricardo Perez-Enriquez. This internship was focused on genetic molecular analysis, disease and pedigree identification in shrimp. In 2016, he started his M. Sc. at UNESP, in the animal breeding and genetics post-graduate program, with a scholarship from CAPES, under the supervision of Dr. Roberto Carneiro finishing it in 2018. In the same year he started his PhD study at the same program and supervisor.

## DEDICATION

I dedicate this dissertation to my family. My mother Silvana, my father Baltasar and my sister Ana Carolina. To my grandmother Maria and my fiancée Rachel.

## ACKNOWLEDGMENTS

First of all, I thank to God for all blesses in my life.

I would like to thank Dr. Roberto Carneiro for all the teachings and advices. I thank for all patience, kindness, and for being an example as animal scientist.

I thank to Dra. Sandra Aidar de Queiroz, Dra. Lucia Galvão de Albuquerque and Dr. Danísio Prado Munari, animal breeding and statistics professors on my undergraduate and post-graduate period. Thank you for all the knowledge and incentive to study animal breeding.

I would like to express my gratitude to all members of the internal and external examining committees: Dr. Danísio Prado Munari, Dr. Henrique Nunes de Oliveira and Dr. Ricardo Vieira Ventura for spending their valuable time and attention with this dissertation and for their priceless suggestions and contributions.

I wish to thank MSc Thaise Pinto de Melo for all help and patience. I learned so much with you. Thank you for your valuable cooperation.

I thank to Dr. Haroldo Henrique de Rezende Neves in name of GenSys Consultores Associados for all the support and effort in developing this dissertation.

I am really grateful to the Genetics and Animal Breeding post graduate program and Animal Science undergraduate course at FCAV- UNESP Jaboticabal. Thank you for all knowledge presented to me since 2010.

I thank the financial support from Coordenação de Aperfeiçoamento de Pessoal de nível Superior (CAPES).

I wish to thank my friends from the shrimp farming sector at CAUNESP. In special to Rafael, Caio and Roberto. Thank you for the friendship and advices all these years.

I thank to my friends of the animal science department: Grazy, Henrique, Giovana, Laiza, Thaise, Daiane, Angel, Helsi, André and Andrés.

I thank to my friends: Leonardo (Jamal), Guilherme (Guizim), Kayo (Diabético), Jhonatan (Jhow), Vitor (Vitão), José (Zé-Mayer), Camilotti, João, Matheus (Bolacha). Thank you for all memorable moments and friendship.

To my fiancée Rachel for all loving support and for being my partner in the difficult and good times.

## CONTENTS

<b>ABSTRACT</b> .....	<b>ii</b>
<b>RESUMO</b> .....	<b>iv</b>
<b>Literature Review</b> .....	<b>1</b>
<b>Introduction</b> .....	<b>5</b>
<b>Material and Methods</b> .....	<b>7</b>
<b>Simulated scenarios</b> .....	<b>7</b>
<b>Population structure</b> .....	<b>7</b>
<b>Genome simulation</b> .....	<b>8</b>
<b>GWAS analyses</b> .....	<b>9</b>
<b>Comparison criteria</b> .....	<b>11</b>
<b>Results and Discussion</b> .....	<b>11</b>
<b>Conclusions</b> .....	<b>19</b>
<b>References</b> .....	<b>20</b>



# POWER OF QTL MAPPING OF DIFFERENT GENOME-WIDE ASSOCIATION METHODS FOR TRAITS UNDER DIFFERENT GENETIC STRUCTURES: A SIMULATION STUDY

## ABSTRACT

The complexity of the traits that can present different genetic structures, such as polygenic or affected by genes of major effect, in addition to different heritabilities, among other factors, make the detection of QTLs challenging. Several methods have been employed with the purpose of performing genome wide association studies (GWAS), aiming the mapping of QTL. The single-step weighted GBLUP (wssGBLUP) method, for example, is an alternative to GWAS, which allows the simultaneous use of genotypic, pedigree and phenotypic information, even from non-genotyped animals. Bayesian methods are also used to perform GWAS, starting from the basic premise that the observed variance can vary at each locus with a specific priori distribution. The objective of the present study was to evaluate, through simulation, which methods, among the evaluated ones, more assist in the identification of QTLs for polygenic and major gene affected traits, presenting different heritabilities. We used the following methods: wssGBLUP, with or without additional phenotypic information from non-genotyped animals and two different weights for markers, where  $w_1$  represented the same weight ( $w_1=1$ ) and  $w_2$  the weight calculated according to the previous iteration process ( $w_1$ ); Bayes C, assuming two values for  $\pi$  ( $\pi = 0.99$  and  $\pi = 0.999$ ), where  $\pi$  is the proportion of SNPs not included in the model, and Bayesian LASSO. The results showed that for polygenic scenarios the detection power is lower and the additional use of phenotypes from non-genotyped animals may help in the detection, yet with low intensity. For scenarios with major effect, there was greater power in the detection of QTL by all different methods with slighter superior performance for the Bayes C method. However, the inclusion of additional phenotypic information caused bias in the estimates and harmed the performance of the wssGBLUP in the presence of major QTL. The increase in heritability for both structures improved the performance of the methods and the power of mapping. The most suitable method for the

detection of QTL is dependent on the genetic structure and the heritability of the trait, and there is not a superior method for all scenarios.

Keywords: GWAS; Genetic Structure; wssGBLUP; Bayesian Methods.

## PODER DE MAPEAR QTL DE DIFERENTES MÉTODOS DE ASSOCIAÇÃO GENÔMICA AMPLA PARA CARACTERÍSTICAS COM DIFERENTES ESTRUTURAS GENÉTICAS: ESTUDO DE SIMULAÇÃO

### RESUMO

A complexidade das características que podem apresentar diferentes estruturas de ação gênica como, por exemplo, poligênicas ou afetadas por genes de efeito maior, aliado a diferentes herdabilidades, entre outros fatores, tornam a detecção de QTLs desafiadora. Diversos métodos têm sido empregados com o intuito de realizar estudos de associação ampla do genoma (GWAS), objetivando o mapeamento de QTL. A metodologia weighted single-step GBLUP (wssGBLUP), por exemplo, é uma alternativa para a realização de GWAS, que permite o uso simultâneo de informações genóticas, de pedigree e fenóticas, mesmo de animais não genotipados. Métodos Bayesianos também são utilizados para a realização de GWAS, partindo da premissa básica de que a variância observada pode variar em cada locus em uma distribuição a priori específica. O objetivo do presente estudo foi avaliar, por meio de simulações, quais métodos, dentre os avaliados, mais auxiliaria na identificação de QTLs para características poligênicas e afetadas por genes de efeito maior, apresentando diferentes herdabilidades. Utilizamos os métodos: wssGBLUP, com a inclusão ou não de informação adicional fenotípica de animais não genotipados e dois distintos ponderadores para os marcadores, onde  $w_1$  representou a mesma ponderação ( $w_1=1$ ) e  $w_2$  a ponderação calculada de acordo com o processo de iteração anterior ( $w_1$ ); Bayes C, assumindo dois valores para  $\pi$  ( $\pi=0.99$  and  $\pi=0.999$ ), onde  $\pi$  é a proporção de SNPs não incluída no modelo, além do LASSO Bayesiano. Os resultados mostraram que para cenários poligênicos o poder de detecção é menor e o uso adicional de fenótipos de animais não genotipados pode ajudar na detecção, ainda que com pouca intensidade. Para cenários com característica sob efeito maior, houve maior poder na detecção de QTL pelos diferentes métodos em comparação aos cenários poligênicos com destaque para a leve vantagem do método Bayes C. A inclusão de informação fenotípica adicional, entretanto, causou viés nas estimativas e atrapalhou o desempenho do wssGBLUP na presença de QTL com efeito maior. O aumento da

herdabilidade para ambas as estruturas melhorou o desempenho dos métodos e o poder de mapeamento. O método mais adequado para a detecção de QTL depende da estrutura genética e da herdabilidade da característica, não existindo um método que seja superior para todos os cenários.

Palavras-chave: GWAS; Estrutura Genética; wssGBLUP; Métodos Bayesianos.

## Literature Review

Several strategies have been adopted to increase productivity and profitability in agricultural systems due to human population growth and environmental resources limitation. Among them, we highlight the application of selective breeding which has been extensively used in animal breeding (Bourdon et al., 2000). Statistical methods based on quantitative genetics and, more recently, molecular approaches are tools often used to help in this task.

Through the advance of technology, such as high density marker panels and bioinformatical and statistical techniques, the use of genomic information was enabled to estimate more accurately the breeding values of animals showing positive results in a relative short-term period as experienced by the dairy cattle sector (Hayes et al., 2009). The understanding of how the target trait is expressed and the detection of possible genomic regions related to it has been an important subject of investigation. Genome-wide association studies (GWAS) were developed aiming to obtain statistical associations between the target trait and markers (Goddard and Hayes, 2009).

However, several traits with economic importance are affected by a large number of genes with small effect hardening GWAS. In addition, different factors may affect the power of detecting quantitative trait loci (QTL), namely: linkage disequilibrium between markers and QTL, trait's heritability, availability of genotypic information and genetic structure, among others (Van den Berg et al., 2013).

Different methods and approaches are used in order to perform GWAS. The Bayesian methods were primary developed to genomic selection purposes and then fitted to GWAS. They usually assume that the variance may fluctuate at each locus with a specific prior distribution. Meuwissen et al. (2001) proposed two Bayesian methods to estimate marker effects, one based on the inclusion of all markers in the model using a scaled  $t$  as prior distribution and a second one, with a fixed proportion of markers included in the model that leads to a posterior distribution of marker effects having higher density at 0, namely Bayes A and Bayes B, respectively. Gianola et al. (2009) observed some drawbacks on these methods related to the prior hyperparameter impacts on shrinkage of SNPs effects and the prior probability  $\pi$ , that a SNP has zero effect is considered as known. Habier et al. (2011) developed

two alternative methods that could overcome these problems. Trying to reduce the scale parameter influence, Bayes C was developed considering the same variance for all SNPs, and Bayes D, with a priori scale parameter as unknown. Another point exploited was the estimation of  $\pi$ , the proportion of SNPs with zero effect. In Bayes A and Bayes B,  $\pi$  was considered as known, thus estimating it directly from the data could improve the results, once  $\pi$  may influence the shrinkage of SNP effects. These modifications originated Bayes C $\pi$  and Bayes D $\pi$  methods. Van den Berg et al. (2013) compared the Bayes C $\pi$  and Bayes C method in a GWAS study and found that fixing the  $\pi$  value is more suitable to QTL detection, instead of calculating it from the data, especially for polygenic low heritable traits.

The Bayesian least absolute shrinkage and selection operator (LASSO) method has also been used in genomic prediction. Tibshirani (1996) proposed it using the sum of the absolute values of the regression coefficients as a penalty in regression models, to simultaneously produce variable selection and shrinkage of coefficients. This operator has the desirable feature of including in the model only a subset of explanatory variables, setting to zero those that have nil effects. De los Campos et al. (2009) modified and extended the LASSO to accommodate pedigree information and marker data into a single model in the context of QTL analysis. Waldman et al. (2013) using real and simulated data showed that LASSO may be an important tool in GWAS, although some issues can possibly affect its performance towards specific scenarios, i.e. most of LASSO methods used in GWAS consider that the sample members are unrelated to each other, which may not be true as often genetic studies enroll multiple members of families (Papachristou et al., 2016).

An alternative approach to perform GWAS is the genomic best linear unbiased prediction (GBLUP) which was firstly developed using multiple step procedures to predict genomic estimated breeding values (GEBV) in genome selection (VanRaden 2008). A limitation for this method is that it uses phenotypes or pseudo-phenotypes only from genotyped animals. To overcome this drawback, a single-step GBLUP method (ssGBLUP) was proposed allowing using all available phenotypic, pedigree and genotypic information simultaneously (Legarra et al., 2009; Christensen and Lund 2010). The method integrates the **G** matrix (genomic relationship) and the **A** matrix (pedigree relationship) into an **H** matrix. The use of ssGBLUP has proven to be successful in studies for different species – beef cattle (Lourenco et al., 2015), dairy cattle (VanRaden 2012), pigs (Forni et al., 2011) and

chicken (Chen et al., 2011; Fragomeni et al., 2014). For GWAS, the ssGBLUP has also been used (ssGWAS). Dikmen et al. (2013) performed ssGWAS for rectal temperature in Holstein Cattle in order to detect association between markers and genes with major effects for this trait. Tiezzi et al. (2015) also used ssGWAS trying to find QTL associated with clinical mastitis in first parity of U.S. Holstein cows.

However, a remarkable drawback for ssGWAS is the prior assumption of equal variance for all markers, which is not necessarily true when the target trait is affected by markers with more pronounced effects than others. Wang et al. (2012) suggested to weight markers variance according to their importance for the trait, i.e., to perform the ssGBLUP principles with the possibility of using all phenotypic (even from non-genotyped animals), genotypic and pedigree information available adding optimal weights for markers variance corresponding to their contribution to the trait. This method, known as weighted ssGBLUP (wssGBLUP), is quite useful for some scenarios such as: plenty more phenotypic than genotypic information, a common condition in real datasets; and complex models (e.g. non-linear; multiple traits). Studies comparing GWAS methods showed that the most suitable method depends on the amount of genotypic and phenotypic data available and the proportion of markers linked to the genetic variance (Lourenço et al., 2014; Wang et al., 2014; Melo et al., 2016).

GWAS have increased significantly with multiple QTL identified to different traits and species. However, a short list of these QTL were validated or reproduced by other studies demanding attention for how the results should be interpreted (Fragomeni et al., 2014). The data simulation is an important tool to previously know important genomic regions and evaluate different methods. For example, Vitezica et al. (2011) compared multiple and single-step procedures considering the ability to predict the GEBV. Wang et al. (2012) also tested different methods and levels of information inclusion observing the accuracy of prediction. Van den Berg et al. (2013) evaluated two methods regarding different scenarios and availability of information. Melo et al. (2016) verified the power of QTL detection under different linkage disequilibrium and using different methods.

The QTL detection remains a challenging task in real data. For instance, despite the decrease in genotyping costs, methods that use less genotypic information might be very interesting, since they do not compromise efficiency and accuracy to the predictions. In this sense, the present study was developed aiming

to: 1) assess which method present better power for QTL mapping in scenarios with different genetic architectures and trait heritability and 2) evaluate the benefit of inclusion of additional phenotypic information from non-genotyped animals in the QTL mapping.



## Introduction

From the evolution of new technologies such as, high density marker panels and advances in bioinformatics, the mechanisms that affect the expression of traits with economic relevance have been more understood in several species (Goddard and Hayes 2009). Genome wide association studies (GWAS) are being adopted in order to assist in the identification of QTL (quantitative trait loci), allowing the detection of statistical associations between important traits and available markers. The identification of QTL may allow obtaining more accurate predictions of the genetic merit and, as a result, improve the genetic progress (Meuwissen et al., 2001).

Number of markers related to the expression of a trait, genetic structure, linkage disequilibrium between markers and QTL, trait's heritability, and the availability of phenotypic and genotypic information, are important issues that should be analyzed in the QTL mapping (Van den Berg et al., 2013). In terms of genetic structure, two main situations that may present difficulties to GWAS are either that a trait is controlled by common variants under small phenotypic effect or in contrast, several uncommon variants, each with a large effect on the phenotype. In both cases the causative mutation may be clustered in a single gene or a small number of genes (major effect), or across several genes (polygenic). For this reason, statistical methods to correct associate the markers with these causal mutations are needed in order to achieve better accuracy in the mapping.

Several methods are available to perform GWAS but there is no agreement whether there is a most suitable method. An approach widely used in GWAS is the multiple regression Bayesian models (Fernando and Garrick, 2013). These methods are able to produce shrinkage and/or variable selection, including all markers simultaneously, assuming different prior distributions. The Bayes A and Bayes B (Meuwissen et al., 2001), Bayes C and Bayes C $\pi$  (Habier et al., 2011) and Bayesian LASSO (Tibshirani 1996, De los Campos et al., 2009) are examples of Bayesian methods often used on GWAS.

The Bayes B method has a fixed proportion of markers included in the model that leads to a posterior distribution of marker effects having higher density at 0. Gianola et al. (2009) suggested modifications on this method arguing that even with marker-specific variances on their models, the shrinkage of effects will still depend

strongly on the prior distribution. In this sense, Habier et al. (2011) proposed the Bayes C $\pi$  method. This method has the prior assumption that marker effects have identical and independent mixture distributions and  $\pi$  (the proportion of SNPs with zero effect) is treated as unknown with a uniform prior. Van den Berg et al. (2013) used this method in a simulation study and concluded that when a small number of records are available and/or several QTL are affecting a trait with low heritability, to apply a fixed value for  $\pi$  is more efficient for QTL mapping than estimate it direct from data. Melo et al. (2016) also used the Bayes C $\pi$  method, with two different values for  $\pi$ , and found that applying a higher shrinkage did not necessarily reflect better QTL detection.

The Bayesian least absolute shrinkage and selection operator (LASSO) method (Tibshirane, 1996) has also been used in GWAS. Its main desirable feature is including just the explanatory fraction of SNPs and setting to zero SNPs with null effect. De los Campos et al. (2009) modified and extended the LASSO to accommodate pedigree information and marker data into a single model in the context of QTL analysis. Waldmann et al. (2013) applied LASSO in real and simulated data showing that despite of small number of true QTL were detected none false positive was found.

The single-step GBLUP (ssGBLUP) is another method used in GWAS (Misztal et al., 2009; Christensen and Lund, 2010). The main feature of this method is to use simultaneously all available pedigree, phenotypic and genotypic information. This is possible due to an **H** matrix that comprises information from **G** matrix (marker information) and **A** matrix (pedigree information) (Legarra et al., 2009). A possible drawback for this method is that equal variances are assumed for the markers, which is not necessarily true regarding traits with larger QTL effects. In this sense, Wang et al. (2012) proposed a method in which the inclusion of all sources of information were possible as well as a weighting the marker according to their importance for the trait (wssGBLUP – weighted ssGBLUP). Indeed, this method has been intensively applied in GWA studies due to the low availability of genotypic information which may be considered insecure when using other methods (Wang et al., 2012; Wang et al., 2014; Silva et al., 2017). In a simulation study, Melo et al. (2016) observed that wssGBLUP exhibited better or similar results than Bayes C $\pi$  in QTL detection. The authors also noted that the use of phenotypic information from non-genotyped animals, in complement to the information of genotyped animals

assisted the QTL detection. However, the authors compared the methods just for a polygenic trait with low heritability. It is unclear in which extension the additional phenotypic information from non-genotyped animals would improve the power of QTL mapping.

The aim of this study was to evaluate, through simulation, the performance of different genome-wide association methods for QTL mapping, considering traits under different genetic structures and heritabilities. We also investigated the impact of additional phenotypic information from non-genotyped animals on QTL mapping for different scenarios.

## **Material and Methods**

### **Simulated scenarios**

The simulation was performed using the QMsim v.1.10 software (Sargolzaei and Schenkel, 2013). Four different scenarios (SI to SIV) were simulated being comprised by hypothetical traits of low (0.14) and moderate (0.35) heritabilities, under polygenic and major gene effects, as described below:

SI:  $h^2=0.14$ , polygenic effect;

SII:  $h^2=0.35$ , polygenic effect;

SIII:  $h^2=0.14$ , major gene effect;

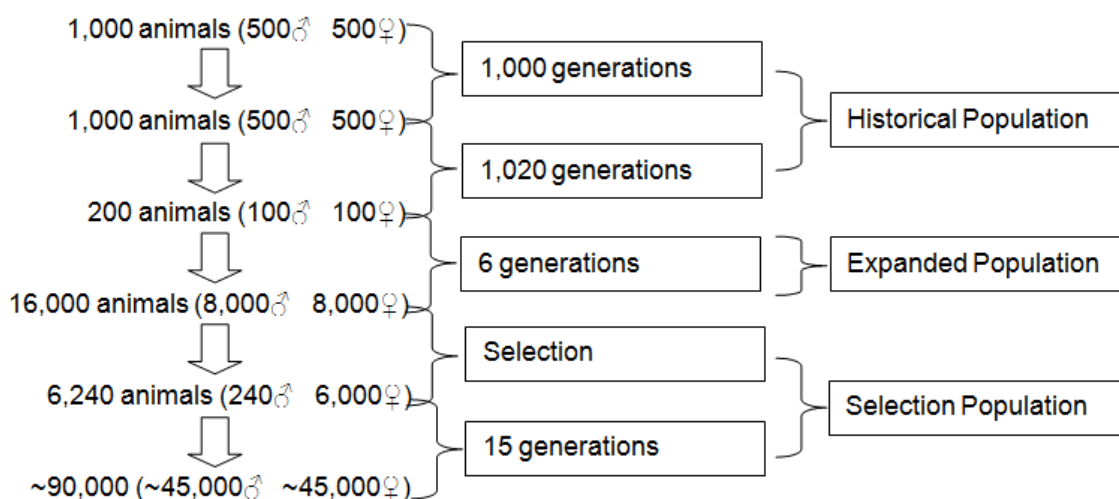
SIV:  $h^2=0.35$ , major gene effect.

To improve the inference power, 20 replicates of each scenario were simulated.

### **Population structure**

Following Melo et al. (2016), a historical population was simulated from generation zero until generation 1000, with a constant size of 1,000 animals (**Figure 1**). Later, from generation 1001 until generation 2020, the population was gradually reduced (from 1000 to 200 animals), producing a “bottleneck” effect and, as a result, genetic drift and linkage disequilibrium compatible with that reported in real populations (O’Brien et al., 2014). The remaining 200 animals, from the last historical generation, were selected to expand the population. In the expansion process, it was assumed random union of gametes in the matings, absence of selection and exponential growth in the number of females, with a 100% replacement rate during

each generation and a mean of five animals per female. Six generations of expansion were generated, resulting in 16,000 animals (8,000 females). After the expanded phase, 240 males and 6,000 females were randomly selected and mated, constituting the founders of the selection population, simulated for 15 generations. At each generation of the selection population, males and females selected were randomly mated, generating a single progeny with equal probability of being male or female. The replacement rate for males and females were kept in 20% and the selection criteria was based on the expected breeding value. The 15 generations of selection population resulted in phenotypic information of 90,000 animals. For the genotypic data, 2,000 animals from the last three generations of selection population were randomly selected and had their genotypes used in GWAS.



**Figure 1. Population structure simulated in all scenarios.**

### Genome simulation

It was assumed that the QTL explained 100% of the genetic variance. The genome had approximately 2,333 cM of length, 735,293 markers and 7,000 QTL. The number of marker and QTL ranged per chromosome from 12,931 to 46,495 and 121 to 438, respectively, being randomly distributed over 29 autosomes. All markers were bi-allelic mimicking the commercial panels available. For the QTL, the number of alleles ranged randomly from two to four. A mutation rate of  $10^{-4}$  for markers and QTL were admitted in the historical population. A total of 335,000 markers (MAF  $\geq$  0.2) and 1,000 QTL were selected from the last historical generation composing the genotypic dataset for the selected population. The average distance among markers

was 0.007 cM. For the polygenic traits, the QTL effects were sampled from a gamma distribution with shape 0.4 (Hayes and Goddard, 2001).

For the major gene effect traits, a polygenic trait was simulated through QMSim as described previously. Then, five QTL were randomly chose each one explaining 7% of additive genetic variance (35% major effect). This effect was produced in the first generation of the selection population replacing the pure polygenic effect by the major effect structure and then passed through all subsequent generations. The remaining 995 QTL were randomly sampled using a gamma distribution with shape 0.4 accounting for the 65% of the variance such as in the polygenic scenario. These procedures were performed using specific routines developed in the R software, version 3.4 (2017).

The phenotypes were the sum of QTL plus an error following normal distribution with mean zero and variance of 0.86 and 0.65, for the low and moderate heritability traits, respectively. The phenotypic variance of traits was standardised in one.

### **GWAS analyses**

Three different methods were compared in terms of QTL detection, namely: wssGBLUP (Wang et al., 2012), Bayes C (Habier et al., 2011) and Bayesian LASSO (Tibshirani, 1996; De los Campos et al., 2009).

The wssGBLUP method was adopted based on the model:  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_a\mathbf{a} + \mathbf{e}$ , where  $\mathbf{y}$  is the vector of phenotypic observations,  $\mathbf{X}$  is the incidence matrix that relates the phenotypes to the fixed effects,  $\boldsymbol{\beta}$  is the vector of fixed effects (overall mean),  $\mathbf{Z}_a$  is the incidence matrix that relates the animals to the phenotypes,  $\mathbf{a}$  is the vector of direct additive genetic effects and  $\mathbf{e}$  is the vector of residuals. The covariance between  $\mathbf{a}$  and  $\mathbf{e}$  was admitted as zero and their variances  $\mathbf{H}\sigma_a^2$  and  $\mathbf{I}\sigma_e^2$ , respectively, where  $\sigma_a^2$  and  $\sigma_e^2$  are the direct additive and residual variances, respectively.  $\mathbf{H}$  is the matrix that combines pedigree and genomic information (Aguilar et al., 2010) and  $\mathbf{I}$  is an identity matrix.

The solutions for the SNP effects ( $\hat{\mathbf{u}}$ ) were obtained according to VanRaden et al. (2009) and Strandén and Garrick (2009), as:  $\hat{\mathbf{u}} = \mathbf{DZ}'[\mathbf{ZDZ}'\mathbf{J}^{-1}\hat{\mathbf{a}}_g]$ , where  $\mathbf{D}$  is the diagonal matrix with weighting factors for the SNPs,  $\mathbf{Z}$  is a matrix that relates the genotypes in each locus, and  $\hat{\mathbf{a}}_g$  is the vector of breeding values predicted for the genotyped animals. The  $\mathbf{D}$  matrix, SNP effects and breeding values were iteratively

calculated according to Wang et al. (2012). Two iteration processes (**w1** and **w2**) were performed for each scenario. The **w1** represented the situation that the same weight (**w1**=1) were attributed for all SNPs. For the **w2** iteration, **d<sub>i</sub>** was calculated according to:  $d_i = \hat{u}_i^2 p_i (1 - p_i)$ , where  $\hat{u}_i$  is the allele substitution effect of the  $i^{th}$  marker estimated from the previous iteration (**w1**), and  $p_i$  is the allele frequency of the second allele of the  $i^{th}$  marker. As a result, a greater shrinkage was applied in **w2** for the SNPs explaining lower variance and, consequently, an increasing proportion of variance was explained by the remaining markers.

In addition to **w1** and **w2** iteration processes as weighting, analyses including all phenotypic information available (even from non-genotyped animals) (**C**) and using only phenotypic information from genotyped animals (**S**) were applied. Thus, four different analyses using wssGBLUP were performed for each scenario. These analyses were performed using the BLUPF90 family programs (Misztal et al., 2012).

The Bayes C method was based on the model:  $y = \mathbf{1}\mu + \sum_{i=1}^n \mathbf{g}_i \mathbf{b}_i \delta_i + \mathbf{e}$ , where the vectors **y** and **e**, are the vectors previously described, **1** is a vector of ones,  $\mu$  is the overall mean,  $\mathbf{g}_i$  is the vector of genotypes for the  $i^{th}$  SNP,  $\mathbf{b}_i$  is the allelic substitution effect,  $\delta_i$  is an indicator variable (0,1) sampled from a binomial distribution with  $n$  and  $\pi$  parameters, which  $n$  is the number of SNPs and  $\pi$  is the proportion of SNPs not included in the model. Two  $\pi$  values were tested: almost fixed to 0.99 (Bayes C $\pi$ I) and 0.999 (Bayes C $\pi$ II), assuming a prior beta distribution with  $\alpha = 10^8$  and  $\beta = 10^{10}$  or  $\beta = 10^{11}$ , respectively. A scaled inverse chis-squared prior distribution was assumed for SNP variance effects ( $\sigma_g^2$ ) and residual variance ( $\sigma_e^2$ ). The Markov chain and Monte Carlo algorithm (MCMC) were applied to the Bayes C method performed on GS3 software (Legarra et al., 2014). A chain with 550,000 iterations, burn-in period of 50,000, and thinning interval of 50 iterations were adopted for this procedure.

The Bayesian LASSO method was implemented in a linear mixed model assuming an exponential prior distribution for variances of SNP marker effects. The LASSO prior distribution for an individual SNP ( $\mathbf{a}_i$ ) follows:  $P(\mathbf{a}_i/\tau^2) \sim N(0, \tau^2)$  and  $Pr(\tau^2) = 0.5\lambda^2 \exp(-\lambda^2/\tau^2)$ . This parameterization process means that individual variances for each SNP (i.e.  $\tau^2$ ) are estimated, according to a regularization parameter ( $\lambda$ ), which was estimated by using a prior gamma distribution restricted

between 0 and  $10^7$ . The MCMC features were the same performed in the Bayes C method.

To summarize, for each scenario 7 different analyses were performed: wssGBLUP assuming two weights ( $w_1$  and  $w_2$ ), with different levels of information inclusion: all phenotypic information available (**C**) and phenotypic information only from genotyped animals (**S**); Bayes C assuming two different values for  $\pi$  ( $\pi=0.99$  and  $\pi=0.999$ ) and the Bayesian LASSO, as described above. As 4 different scenarios and 20 replicates per scenario were simulated, a total of 560 (4x20x7) GWAS analyses were performed.

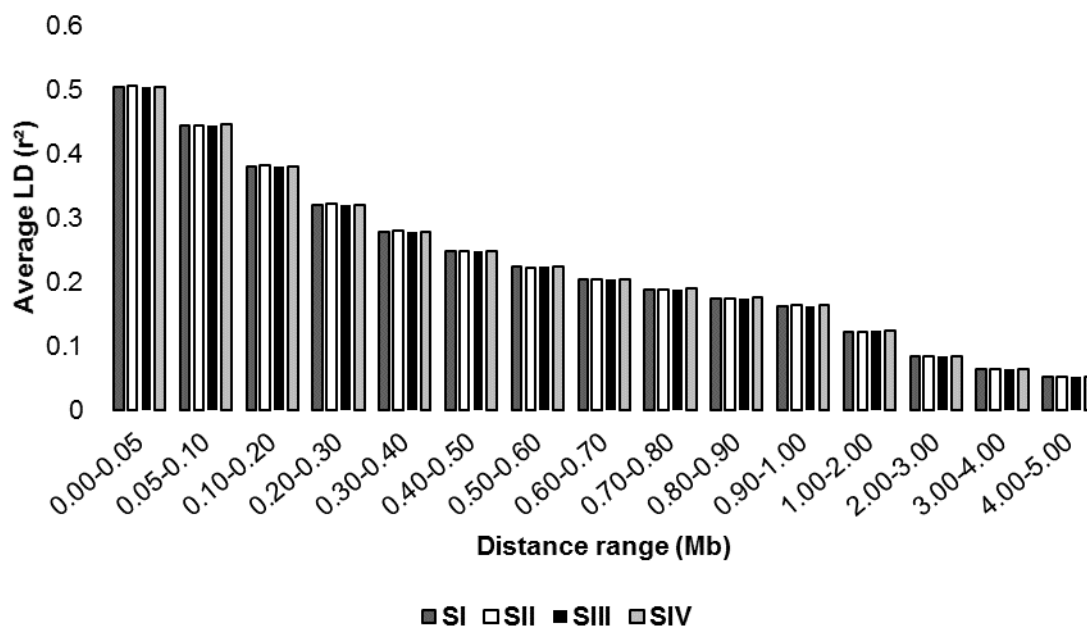
### Comparison criteria

The linkage disequilibrium decay (LD) between any two loci was surveyed and compared to similar study for all scenarios using the  $r^2$  parameter which is supplied by a QMSim feature. The following criteria was used to compare the methods: number of QTL that explained more than 1% of the genetic variance (**NtopQ**); number of windows (1 Mb) with greater proportion of variance explained by the markers (**topM**); sum of all genetic variance explained by the NtopQ (**P\_topQ**) and superior marker windows (**P\_topM**); maximum percentage explained by a topQTL (**P\_1<sup>st</sup>Q**) and for the superior marker window (**P\_1<sup>st</sup>M**); number of true QTL identified (**NtrueQ**), i.e. the number of **topQ** identified by a **topM** no further than 1Mb from the true QTL position. The **P\_1<sup>st</sup>True** value indicates the number of times that a method was the best in 20 replicates i.e. which method(s) showed the highest ability to detect QTL in each replicate. In case of draw among methods, one point was attributed to each method, thus the sum of all methods surplus 100%.

### Results and Discussion

The average linkage disequilibrium decays over the 20 replicates for each scenario are shown in **Figure 2**. As expected, no expressive variation among the scenarios was found, since they were simulated through the same population structure. The same LD pattern was observed by Melo et al. (2016) and Pérez O'Brien et al. (2014) for a simulated and real population, respectively, highlighting the adequacy of the simulated populations.

**Figure 3** shows how the QTL effect was distributed on the genome post-simulation. For the polygenic trait with lower and higher heritability, SI and SII, respectively, it is possible to notice the stronger peaks in SII due to higher proportion of additive genetic variance. For SIII and SIV, we were able to produce the desired major effect (5 QTL summing up for 35% of total genetic variance). Often these 5 QTL presented in the major effect structure scenarios were the only topQ (greater than 1% of total variance) found, however, in some replicates the “polygenic part” exhibited topQs as well.



**Figure 2. Linkage disequilibrium (LD) decay of all scenarios (SI to SIV) simulated. Average LD, expressed in  $r^2$ , according to varying distances between markers (Mb).**

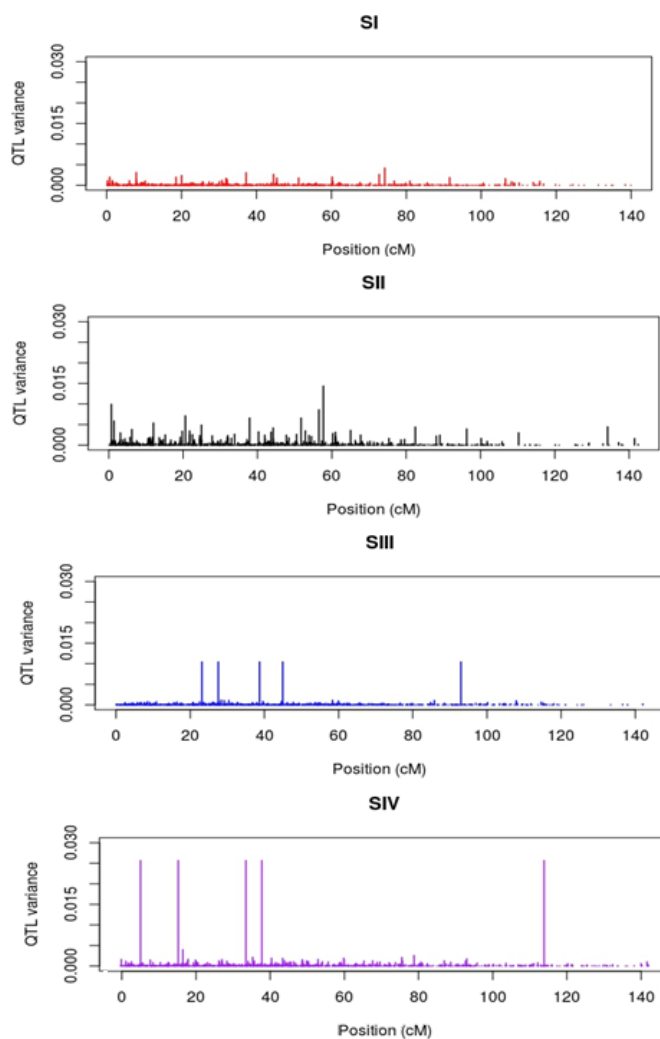
As observed by Melo et al. (2016), in the polygenic scenario under low heritability, independent of the statistical method adopted, the QTL detection was poor. On average, 17.55 QTL were simulated over the 20 replicates and only 2.6 (14.8%) to 3.75 (21.4%) QTL were detected, depending on the method adopted (**Table1**). This result is in agreement with Van den Berg et al. (2013) who highlighted the difficulty of QTL mapping for low heritable polygenic traits.

Although the comparison among methods was difficult due to high standard deviations on the means, the relative number of best detection ( $P_{1^{st}True}$ ) helped to



show that CW1 model presented better results for QTL mapping as the best method in 45% of replicates. The slight superiority of CW1 may be explained by the wssGBLUP model premises which fit to an infinitesimal model, i.e., many markers presenting small effects.

The effect of additional phenotypic information helped slightly the QTL detection. For CW1, on average, the use of this extra information, allowed the detection of 21.36% (3.75 out of 17.55) NtopQ, compared to 15.10% (2.65 out of 17.55) when this information was ignored (SW1). These results are in agreement with Melo et al. (2016) with 17.36% and 11.98% also for w1, for inclusion and non-inclusion of additional information, respectively. The same pattern was observed for w2 in this scenario.



**Figure 3. Examples of QTL variance and distribution for random replicates simulated for all scenarios present in the last selection generation. Position in the x-axis refers to position of the QTL in cM irrespective of the chromosome.**

Using a stronger shrinkage on Bayes C (Bayes C $\pi$ II) resulted in greater proportion of the genetic variance being captured by the leading SNPs (P\_topM). However, this did not reflect in greater power to QTL detection, when compared to a weaker shrinkage (Bayes C $\pi$ I). Similarly, using w2 in the wssGBLUP method increased the genetic variance captured by the leading SNPs, but the mean NtrueQ were similar between w1 and w2 or superior for w1 (Table 1: contrasts SW1xSW2 and CW1xCW2).

**Table 1. Mean (SD) for QTL and marker statistics<sup>1</sup> using Bayes C, Bayesian LASSO and weighted single step GBLUP for SI ( $h^2=0.14$ , polygenic effect), over 20 replicates.**

Method <sup>2</sup>	P_topM (%)	P_1 <sup>st</sup> M (%)	NtrueQ	P_1 <sup>st</sup> True (%) <sup>3</sup>
Bayes C $\pi$ I	7.26 (1.09)	0.87 (0.26)	2.95 (2.01)	15.00%
Bayes C $\pi$ II	37.59 (12.56)	9.96 (7.80)	2.95 (2.35)	15.00%
LASSO	5.25 (0.63)	0.50 (0.09)	2.60 (1.76)	20.00%
CW1	5.17 (0.06)	0.50 (0.09)	3.75 (1.74)	45.00%
CW2	16.84 (2.48)	2.09 (0.58)	3.00 (2.00)	30.00%
SW1	5.42 (0.68)	0.52 (0.10)	2.65 (1.60)	10.00%
SW2	21.23 (4.43)	3.22 (1.53)	2.85 (1.95)	10.00%
True Values	P_topQ (%)	P_1 <sup>st</sup> Q (%)	NtopQ	
	31.30 (3.55)	3.96 (1.42)	17.55 (2.01)	

<sup>1</sup> Pvar\_topMRKw (P\_topQ): sum of all genetic variance explained by the NtopQ (P\_topQ) and superior marker windows (P\_topM);

P\_1<sup>st</sup>M (P\_1<sup>st</sup>Q): maximum percentage explained by a topQ (P\_1<sup>st</sup>Q) and for the superior marker window (P\_1<sup>st</sup>M);

NtrueQ (NtopQ): number of true QTL identified (NtrueQ) and number of QTL that explained more than 1% of the genetic variance (NtopQ);

<sup>2</sup> Bayes C $\pi$ I (Bayes C $\pi$ II): Bayes C assuming two different values for  $\pi$ ,  $\pi=0.99$ , (BayesC $\pi$ I) and  $\pi=0.999$ , (Bayes C $\pi$ II);

LASSO: Bayesian LASSO;

CW1 (CW2): wssGBLUP with the inclusion of all phenotypic information, including from non genotyped animals and applying different weights w1 (CW1) and w2 (CW2);

SW1 (SW2): wssGBLUP with the inclusion of phenotypic information only from genotyped animals and applying different weights w1 (SW1) and w2 (SW2);

<sup>3</sup> P\_1<sup>st</sup>True: number of times that each method showed the highest NtrueQ.

For SII, the number of NtopQ, proportion of maximum variance explained by a single QTL and sum of topQ variances were very similar to SI: 30.07%, 3.83% and 17.25, respectively (**Table 2**). It is also possible to observe a greater power of QTL detection for all methods due to heritability increase, compared to SI. For SII, the

QTL detection (NtrueQ) ranged from 31.20% (5.40 out of 17.25) to 36.2% (6.25 out of 17.25). The influence of heritability on QTL detection was also highlighted by Van den Berg et al. (2013), where the number of false positives decreased as the heritability increased.

**Table 2. Mean (SD) for QTL and marker statistics<sup>1</sup> using Bayes C, Bayesian LASSO and weighted single step GBLUP for SII ( $h^2=0.35$ , polygenic effect), over 20 replicates.**

Method <sup>2</sup>	P_topM (%)	P_1 <sup>st</sup> M (%)	NtrueQ	P_1 <sup>st</sup> True (%) <sup>3</sup>
Bayes C $\pi$ I	12.90 (3.87)	2.32 (2.58)	6.11 (3.74)	40.00%
Bayes C $\pi$ II	66.57 (7.93)	22.16 (10.63)	6.25 (3.04)	30.00%
LASSO	5.50 (0.92)	0.57 (0.16)	5.80 (3.37)	20.00%
CW1	5.51 (1.00)	0.57 (0.20)	6.20 (2.97)	25.00%
CW2	16.72 (4.82)	2.61 (1.38)	5.45 (2.35)	15.00%
SW1	5.53 (0.93)	0.57 (0.16)	5.65 (3.39)	20.00%
SW2	17.58 (2.66)	2.55 (1.41)	5.40 (2.95)	0.00%
True	P_topQ (%)	P_1 <sup>st</sup> Q (%)	NtopQ	
Values	30.07 (5.68)	3.83 (1.13)	17.25 (3.70)	

<sup>1</sup> Pvar\_topMRKw (P\_topQ): sum of all genetic variance explained by the NtopQ (P\_topQ) and superior marker windows (P\_topM);

P\_1<sup>st</sup>M (P\_1<sup>st</sup>Q): maximum percentage explained by a topQ (P\_1<sup>st</sup>Q) and for the superior marker window (P\_1<sup>st</sup>M);

NtrueQ (NtopQ): number of true QTL identified (NtrueQ) and number of QTL that explained more than 1% of the genetic variance (NtopQ);

<sup>2</sup> Bayes C $\pi$ I (Bayes C $\pi$ II): Bayes C assuming two different values for  $\pi$ ,  $\pi=0.99$ , (BayesC $\pi$ I) and  $\pi=0.999$ , (Bayes C $\pi$ II);

LASSO: Bayesian LASSO;

CW1 (CW2): wssGBLUP with the inclusion of all phenotypic information, including from non genotyped animals and applying different weights w1 (CW1) and w2 (CW2);

SW1 (SW2): wssGBLUP with the inclusion of phenotypic information only from genotyped animals and applying different weights w1 (SW1) and w2 (SW2);

<sup>3</sup> P\_1<sup>st</sup>True: number of times that each method showed the highest NtrueQ.

Again, Bayes C $\pi$ II, CW2 and SW2 captured greater part of genetic variance in contrast to the other methods. However, Bayes C $\pi$ I and Bayes C $\pi$ II sharply increased this amount of genetic variance captured (56.28 and 56.47%, respectively, more in comparison to the same methods on SI). For CW2 and SW2 the shrinkage did not produce the same effect on the variance captured and it remained practically constant in relation to SI statistics.

In comparison to SI, there was more similarity among the methods regarding NtrueQ. The benefit of additional phenotypic information in wssGBLUP on QTL detection seemed less pronounced. For w1 iteration, 35.94% (6.20 out of 17.25) NtopQ were detected (CW1) while, ignoring this information (SW1) this number reduced only to 32.75% (5.65 out of 17.25). It most likely happened because of the additional phenotype information from related animals has less impact on the accuracy of estimated breeding values, and as a result, less importance on the estimatives of SNP effects, when compared to a lower heritability trait. Similar tendency was found for w2 iteration with 5.45 and 5.40 for CW2 and SW2, respectively.

For SIII (**Table 3**) the results suggested that all methods evaluated had superior power for QTL detection when the trait presents major QTL effect. On average, NtrueQ detected ranged from 31.1% (1.65 out of 5.30) to 53.8% (2.85 out of 5.30). This result was expected due to the major effect simulated (5 QTL explaining 7% of total genetic variance each) concentrating the effect and favouring all methods in the estimation of SNP effects.

For this scenario, the Bayes C method showed results slight superior than the other methods as showed by 75% and 60% of best detection ( $P_{1^{st}True}$ ) for Bayes C $\pi$ I and Bayes C $\pi$ II, respectively. Most likely the assumptions of Bayes C method were better fitted to the genetic structure simulated. Mehrban et al. (2017) comparing methods for genomic selection found that Bayes C exhibits more accuracy for carcass weight in Hanwoo beef cattle, a possible indicative of major gene effects. This had been supported by Lee et al. (2013) through an association study detecting 6 SNP loci for carcass weight with some of these markers accounting for more than 10% of total genetic variance. As for the other scenarios, the increase on shrinkage (Bayes C $\pi$ II x Bayes C $\pi$ I) was not associated to a better QTL detection. This probably happened due to proportionally few genotypes and low heritability trait on this scenario.

Distinct from the polygenic scenarios, the additional phenotypic information tended to reduce the power of QTL detection in SIII (**Table 3**: contrasts CW1xSW1 and CW2xSW2, for the different levels of information included in the wssGBLUP method). Most likely, the extra phenotypic information compromised the QTL detection by introducing “noise” on wssGBLUP estimations. Close relatives not necessarily present the same genotype regarding the major QTL configuration, and

using their phenotypic information without knowing their genotypes may hamper QTL mapping.

Although, exactly 5 major QTL with 7% of total genetic variance were simulated, on average 5.3 QTL were found over 20 replicates. The surplus 0.3 is associated to others QTL that in determined replicates explained more than 1% of the genetic variance being accounted for the total variance explained by all QTL. The milder reduction on the proportion of total genetic variance explained by the major QTL (6.95 instead of 7.00) may indicate that the polygenic part also exhibited NtopQ(5.30, instead of exactly 5) along the 20 replicates.

**Table 3. Mean (SD) for QTL and marker statistics<sup>1</sup> using Bayes C, Bayesian LASSO and weighted single step GBLUP for SIII ( $h^2=0.14$ , major gene effect), over 20 replicates.**

Method <sup>2</sup>	P_topM (%)	P_1 <sup>st</sup> M (%)	NtrueQ	P_1 <sup>st</sup> True (%) <sup>3</sup>
Bayes CπI	5.49 (1.60)	2.00 (1.01)	2.85(1.01)	75.00
Bayes CπII	51.07 (16.66)	29.47 (16.72)	2.65(0.99)	60.00
LASSO	2.75 (0.57)	0.73 (0.24)	2.26(1.24)	30.00
CW1	2.26 (0.62)	0.59 (0.28)	1.90(0.41)	25.00
CW2	9.85 (2.70)	3.03 (1.16)	1.65(1.18)	15.00
SW1	3.48 (2.89)	1.06 (1.42)	2.40(1.27)	40.00
SW2	16.69 (5.97)	6.58 (3.81)	2.45(1.10)	35.00
True	P_topQ (%)	P_1 <sup>st</sup> Q (%)	NtopQ	
Values	35.08 (1.39)	6.95 (0.29)	5.30 (0.66)	

<sup>1</sup> Pvar\_topMRKw (P\_topQ): sum of all genetic variance explained by the NtopQ (P\_topQ) and superior marker windows (P\_topM);

P\_1<sup>st</sup>M (P\_1<sup>st</sup>Q): maximum percentage explained by a topQ (P\_1<sup>st</sup>Q) and for the superior marker window (P\_1<sup>st</sup>M);

NtrueQ (NtopQ): number of true QTL identified (NtrueQ) and number of QTL that explained more than 1% of the genetic variance (NtopQ);

<sup>2</sup> Bayes CπI (Bayes CπII): Bayes C assuming two different values for π, π=0.99, (BayesCπI) and π=0.999, (Bayes CπII);

LASSO: Bayesian LASSO;

CW1 (CW2): wssGBLUP with the inclusion of all phenotypic information, including from non genotyped animals and applying different weights w1 (CW1) and w2 (CW2);

SW1 (SW2): wssGBLUP with the inclusion of phenotypic information only from genotyped animals and applying different weights w1 (SW1) and w2 (SW2);

<sup>3</sup> P\_1<sup>st</sup>True: number of times that each method showed the highest NtrueQ.

The results on **Table 4** shows that all methods evaluated outperformed the QTL detection when the trait is under major effect and higher heritability. Over all replicates, the extension of QTL detection was from 40.37% (2.20 out of 5.45) to 72.48% (3.95 out of 5.45). This result was also expected due to the increase on additive genetic variance.

The Bayes C method showed greater power to QTL detection compared to the other methods. Again, the genetic structure may have favoured the variance of SNPs captured by this method. However, in scenario SIV, the difference between stronger and weak shrinkage, regarding the variance captured by top markers, seemed to be less pronounced than in SIII.

**Table 4. Mean (SD) for QTL and marker statistics using Bayes C, Bayesian LASSO and weighted single step GBLUP for SIV ( $h^2=0.35$ , major gene effect), over 20 replicates.**

Method	P_topM(%)	P_1 <sup>st</sup> M(%)	NtrueQ	P_1 <sup>st</sup> True (%) <sup>3</sup>
Bayes C $\pi$ I	19.52 (12.18)	10.37 (9.68)	3.90 (1.37)	70.00
Bayes C $\pi$ II	69.76 (11.06)	37.40 (16.32)	3.80 (1.06)	55.00
LASSO	3.25 (1.11)	0.84 (0.35)	3.00 (1.63)	30.00
CW1	2.35 (0.72)	0.57 (0.14)	2.20 (1.61)	10.00
CW2	9.18 (4.18)	3.01 (1.57)	2.35 (1.73)	25.00
SW1	3.42 (0.87)	0.88 (0.25)	3.45 (1.50)	25.00
SW2	16.13 (4.59)	6.24 (3.13)	3.95 (1.05)	40.00
True	P_topQ(%)	P_1 <sup>st</sup> Q (%)	NtopQ	
Values	35.87 (1.88)	7.07 (0.31)	5.45 (0.69)	

<sup>1</sup> Pvar\_topMRKw (P\_topQ): sum of all genetic variance explained by the NtopQ (P\_topQ) and superior marker windows (P\_topM);

P\_1<sup>st</sup>M (P\_1<sup>st</sup>Q): maximum percentage explained by a topQ (P\_1<sup>st</sup>Q) and for the superior marker window (P\_1<sup>st</sup>M);

NtrueQ (NtopQ): number of true QTL identified (NtrueQ) and number of QTL that explained more than 1% of the genetic variance (NtopQ);

<sup>2</sup> Bayes C $\pi$ I (Bayes C $\pi$ II): Bayes C assuming two different values for  $\pi$ ,  $\pi=0.99$ , (BayesC $\pi$ I) and  $\pi=0.999$ , (Bayes C $\pi$ II);

LASSO: Bayesian LASSO;

CW1 (CW2): wssGBLUP with the inclusion of all phenotypic information, including from non genotyped animals and applying different weights w1 (CW1) and w2 (CW2);

SW1 (SW2): wssGBLUP with the inclusion of phenotypic information only from genotyped animals and applying different weights w1 (SW1) and w2 (SW2);

<sup>3</sup> P\_1<sup>st</sup>True: number of times that each method showed the highest NtrueQ.

As for SIII, the use of additional phenotypic information compromised wssGBLUP performance in terms of QTL detection. Again, this result is related to the fact that the major QTL effect may vary in the relatives included as extra information. Both w2 iterations (CW2 x SW2) showed more power in the detection of QTL, compared to w1. These results are in agreement with Zhang et al. (2016) study, in which, the use of different weights for SNPs is more effective for traits influenced by few QTL than in polygenic structures resulting in more power in QTL detection, .

The LASSO method presented steady results over all the scenarios in comparison to other methods. However, the influence of increase on the heritability and the shift to the major effect structure resulted in increase on QTL mapping, same pattern observed for the other methods. Its intermediary performance may be explained by the LASSO assumption that this method cannot select more predictor variables than the sample size. This could potentially be a problem in our study that involves several more predictor variables than response variables. However, some studies state that the double exponential prior used by LASSO may present a remarkable advantage by modelling multiple markers at same time. This method is able to distinct trait supplier loci from others that are in high linkage disequilibrium with those loci. (Papachristou et al., 2016; Motyer et al., 2011).

The proportion of times that each method was the best ( $P_{1^{st}True}$ ) summarizes the results obtained on true QTL detection means highlighting: the benefit of using additional phenotypic information (CW1) on scenarios under complex traits (SI); the similarity among methods when the heritability is increased under polygenic structures (SII); the remarkable performance of Bayes C method, mainly with  $\pi$  almost fixed to 0.99, when the trait is under major structures (SIII and SIV). The most suitable method for each scenario relies on further studies regarding, for instance, other genetic structures, varying number of genotyped animals and different population structures. By investigating the methods' strength and weakness, the QTL mapping will become more accurate and helpful.

## Conclusions

The most suitable method to perform GWAS relies on the genetic structure and trait's heritability.

Despite the small differences among methods in the QTL detection, for polygenic traits, the wssGBLUP method seemed to show better results in comparison to the other methods mainly in the low heritability scenario.

For traits with major gene effects, the Bayes C method is expected to present better results, compared to the other methods evaluated.

The use of additional phenotypic information from non genotyped animals implies in better QTL detection in polygenic traits. However, for traits under major gene effect, it may negatively affect the QTL detection.

## References

- Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of dairy science*, 93(2), 743-752.
- Bourdon, R.M. (1999) Understanding animal breeding. 2nd ed. Englewood Cliffs, NJ: Prentice Hall. 538p.
- Chen, C. Y., Misztal, I., Aguilar, I., Legarra, A., Muir, W. M. (2011). Effect of different genomic relationship matrices on accuracy and scale. *Journal of animal science*, 89(9), 2673-2679.
- Christensen, O. F., Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, 42(1), 2.
- De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1), 375-385.
- Dikmen, S., Cole, J. B., Null, D. J., Hansen, P. J. (2013). Genome-wide association mapping for identification of quantitative trait loci for rectal temperature during heat stress in Holstein cattle. *PLoS One*, 8(7).
- Fernando, R. L., Garrick, D. (2013). Bayesian methods applied to GWAS. Genome-wide association studies and genomic prediction, 237-274.
- Forni, S., Aguilar, I., Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics Selection Evolution*, 43(1), 1.



- Fragomeni, B. O., Misztal, I., Lourenco, D. L., Aguilar, I., Okimoto, R., Muir, W. M. (2014). Changes in variance explained by top SNP windows over generations for three traits in broiler chicken. *Frontiers in genetics*, 5.
- Gianola D., De Los Campos G., Hill W.G., Manfredi E., Fernando R.L. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, 183: 347–363.
- Goddard, M. E., Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10(6), 381-391.
- Habier D., Fernando R.L., Kizilkaya K., Garrick, D.J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12:186.
- Hayes, B. E. N., Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution*, 33(3), 1.
- Hayes, B. J., Bowman, P. J., Chamberlain, A. J., Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, 92(2), 433-443.
- Lee, S. H., Choi, B. H., Lim, D., Gondro, C., Cho, Y. M., Dang, C. G., Sharma, A., Jang, G. W., Lee, K. T., Yoon, D., Lee, H. K., Yeon, S. H., Yang, B. S., Kang, H. S., Hong, S. K. (2013). Genome-wide association study identifies major loci for carcass weight on BTA14 in Hanwoo (Korean cattle). *PLoS One*, 8(10).
- Legarra, A., Aguilar, I., Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, v. 92, n. 9, p. 4656-4663.
- Legarra, A.; Ricard, A.; Filangi, O. GS3. Manual. France. [S.l.: s.n.], 2014, 24p. Available in: <[http://snp.toulouse.inra.fr/~alegarra/manualgs3\\_last.pdf](http://snp.toulouse.inra.fr/~alegarra/manualgs3_last.pdf)>. Accessed in Oct. 2017.
- Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Masuda, Y., Aguilar, I., Legarra, A., Bertrand, J. K., Amen, T. S., Wang, L., Moser, D. W., Misztal, I. (2015). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *Journal of animal science*, 93(6), 2653-2662.
- Mehrban, H., Lee, D. H., Moradi, M. H., IlCho, C., Naserkheil, M., Ibáñez-Escriche, N. (2017). Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. *Genetics Selection Evolution*, 49(1), 1.

- Melo, T. P., Takada, L., Baldi, F., Oliveira, H. N., Dias, M. M., Neves, H. H., Schenkel F.S., Albuquerque, L.G., Carneiro, R. (2016). Assessing the value of phenotypic information from non-genotyped animals for QTL mapping of complex traits in real and simulated populations. *BMC genetics*, 17(1), 89.
- Meuwissen, T. H. E., Hayes, B. J., Goddard, M. E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, n. 4, p. 1819-1829.
- Misztal, I. BLUPF90-a flexible mixed model program in Fortran 90. User Manual. Animal and Dairy Science. Athens, GA, USA: University of Georgia; [S.l.: s.n.], 2012. Available in: <http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90.pdf> . Accessed in: nov. 2017.
- Misztal, I., Legarra, A., Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science*, 92(9), 4648-4655.
- Motyer, A. J., McKendry, C., Galbraith, S., Wilson, S. R. (2011). LASSO model selection with post-processing for a genome-wide association study data set. In *BMC proceedings* (Vol. 5, No. 9, p. S24). BioMed Central.
- O'Brien, A. M. P., Mészáros, G., Utsunomiya, Y. T., Sonstegard, T. S., Garcia, J. F., Van Tassell, C. P., Carneiro, R., Silva, M. V. B. Sölkner, J. (2014). Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. *Livestock Science*, 166, 121-132.
- Papachristou, C., Ober, C., Abney, M. (2016). A LASSO penalized regression approach for genome-wide association analyses using related individuals: application to the Genetic Analysis Workshop 19 simulated data. In *BMC proceedings* (Vol. 10, No. 7, p. 53). BioMed Central.
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Sargolzaei, M., Schenkel, F. S. (2013). QMSim: User's Guide [S.l.: s.n.], p.77.
- Silva, R. M. O., Vallejo, R. L., Evenhuis, J. P., Leeds, T. D., Gao, G., Parsons, J. E., Martin, K. E., Lourenco, D. A. L., Palti, Y. (2017). 209 Prospecting genomic regions associated with columnaris disease in two rainbow trout breeding populations. *Journal of Animal Science*, 95(supplement 4), 103-104.

- Stranden, I., Garrick, D. J. (2009). Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of dairy science*, 92(6), 2971-2975.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
- Tiezzi, F., Parker-Gaddis, K. L., Cole, J. B., Clay, J. S., Maltecca, C. (2015). A genome-wide association study for clinical mastitis in first parity US Holstein cows using single-step approach and genomic matrix re-weighting procedure. *PLoS One*, 10(2).
- Van den Berg, I., Fritz, S., Boichard, D. (2013). QTL fine mapping with Bayes C ( $\pi$ ): a simulation study. *Genetics Selection Evolution*, 45(1), 19.
- VanRaden, P. M. (2012). Avoiding bias from genomic pre-selection in converting daughter information across countries. In: INTERNATIONAL BULL EVALUATION SERVICE, 2., n. 45, p. 1-5. Verona, Italy.
- VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of dairy science*, 92(1), 16-24.
- Vitezica, Z. G., Aguilar, I., Misztal, I., Legarra, A. (2011). Bias in genomic predictions for populations under selection. *Genetics Research*, 93(5), 357-366.
- Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics*, 4.
- Wang, H., Misztal, I., Aguilar, I., Legarra, A., Fernando, R. L., Vitezica, Z., Okimoto, R., Wing, T., Hawken, R., Muir, W. M. (2014). Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Frontiers in genetics*, 5.
- Wang, H., Misztal, I., Aguilar, I., Legarra, A., Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genetics Research*, 94(2), 73-83.
- Zhang, X., Lourenco, D., Aguilar, I., Legarra, A., Misztal, I. (2016). Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. *Frontiers in genetics*, 7.