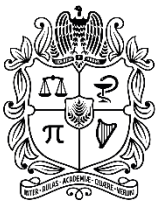


UNIVERSIDAD
NACIONAL
DE COLOMBIA

# Standardization of a methodology for identification and annotation of associations between single nucleotide polymorphisms and highly polygenic traits in ruminants

**Boris Julián Sepúlveda Molina**

Universidad Nacional de Colombia

Facultad de Ingeniería

Departamento de Ingeniería de Sistemas e Industrial

Programa de Maestría en Bioinformática

Bogotá D.C., Colombia

2019

# Standardization of a methodology for identification and annotation of associations between single nucleotide polymorphisms and highly polygenic traits in ruminants

**Boris Julián Sepúlveda Molina**

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

**Magíster en Bioinformática**

Directora:

Ph.D., Liliana López Klein

Línea de Investigación:

Estadística Genómica

Universidad Nacional de Colombia

Facultad de Ingeniería

Departamento de Ingeniería de Sistemas e Industrial

Programa de Maestría en Bioinformática

Bogotá D.C., Colombia

2019

*To my family*

# Acknowledgment


- To Professor Liliana López Kleine for her advice and support in this work.

- To my family for the unconditional support that has given me in all the decisions, I have made in my life.

- To the students, and the academic and administrative staff of the Universidad Nacional de Colombia for their contribution to the maintenance of quality public education, despite the financial and political adversities.

# Abstract

Given the importance of the production of ruminants, it is necessary to investigate the genetic variants associated with the traits of economic interest in these animals, as well as the biology underlying the genotype-phenotype associations. To conduct these associations, a widely used strategy is to perform genome-wide association studies (GWAS). The GWAS must have the support of adequate quality control (QC), to then identify the associations between genetic markers type SNP and phenotypes. Additionally, the biological contextualization of these associations starts from the annotation of the genes close to the associated markers. Currently, there are several tools, including R libraries, to perform these analyses. However, it is necessary to develop a tool that allows unifying the three main steps (QC, GWAS, and annotation) for species other than human. For the above, the present work developed a methodology that unified the three mentioned steps in the R environment. The generated code was submitted for publication and is freely available in the repository https://github.com/bojusemo/Diploid-GWAS. The code was tested in two populations of ruminants, the Colombian Creole Hair Sheep and Simmental cattle. In these populations, the SNPs with low quality were removed, there was no detected population stratification, and no samples were removed for low quality. The SNP OAR26_10469468.1 was associated with the meat tenderness of Colombian Creole hair sheep. This SNP is in the gene TENM3. TENM3 protein has two domains with functions associated with meat tenderness in cattle and pigs. The SNP BovineHD4100012055 was associated with birth weight in Simmental. The closest gene to this SNP is the olfactory receptor 52E8-like, which is a member of the protein family G protein-coupled receptor (GPCR). GPCR has associated with birth weight in humans. Six markers were associated with 305-day milk yield in Simmental. Neither the closest genes of these markers nor their protein domains have been reported as associated with milk production.

**Keywords: genome-wide association studies; single nucleotide polymorphism, annotation, ruminants.**

# Resumen

Dada la importancia que tiene la producción de rumiantes, es necesario investigar las variantes genéticas asociadas a las características de interés comercial de dichos animales, así como la biología subyacente a esas asociaciones genotipo-fenotipo. Para hacer dichas asociaciones, una estrategia ampliamente utilizada es realizar estudios de asociación del genoma completo (GWAS). Los GWAS deben partir de un filtro adecuado de la información de las variables y de los individuos, denominado control de calidad (QC), para luego identificar las asociaciones entre marcadores genéticos tipo SNP y los fenotipos. Por su parte, la contextualización biológica de estas asociaciones parte de la anotación de los genes cercanos a los marcadores asociados. Para realizar estos análisis, actualmente hay varias herramientas, incluidas librerías de R. Sin embargo, falta desarrollar una herramienta que permita unificar los tres principales pasos (QC, GWAS y anotación) para datos de especies distintas al humano en R. Por lo anterior, el presente trabajo desarrolló una metodología que unificó en el entorno de R los tres pasos mencionados. El código generado se sometió a publicación y se encuentran disponibles de manera libre en el repositorio https://github.com/bojusemo/Diploid-GWAS. El código fue probado en dos poblaciones de rumiantes, el Ovino de Pelo Criollo Colombiano y los bovinos Simmental. En estas poblaciones, se eliminaron los SNPs con una baja calidad, no se detectó estratificación poblacional y no se eliminaron muestras por baja calidad. El SNP OAR26_10469468.1 estuvo asociado con la terneza de la carne del Ovino de Pelo Criollo Colombiano. Éste SNP está en el gen TENM3. La proteína TENM3 tiene dos dominios con funciones asociadas con la terneza de la carne en bovinos y porcinos. El SNP BovineHD4100012055 estuvo asociado con el peso al nacimiento de Simmental. El gen más cercano a este SNP es el *olfactory receptor 52E8-like*, que pertenece a la familia de proteínas *G protein-coupled receptor* (GPCR). Se ha reportado asociación entre GPCR y el peso al nacimiento en humanos. Seis marcadores estuvieron asociados a la producción de leche a los 305 días en Simmental. Ni los genes más cercanos a los marcadores, ni los dominios de las proteínas han sido reportados como asociados con la producción de leche.

**Palabras clave: estudios de asociación del genoma completo, polimorfismo de nucleótido simple, anotación, rumiantes.**

# Table of contents

# List of figures

**Pg.**

# List of tables

**Pg.**

# Introduction

One can divide the traits of economic importance in ruminants in those that are determined by few genes, and in those that are highly polygenic, that is, in which a high number of genes determines the effect of genetics. Productive traits as birth weight and milk production, and those of quality of the final product like the meat tenderness are between the most economically interesting ones. Regarding the traits determined by few genes, in many cases, the GWAS have detected SNPs that are in gene regions that have a significant effect on the phenotype and have allowed elucidating their mechanism of action. On the other hand, for the highly polygenic traits, these studies have contributed to predicting the phenotype from the genotype based on the detection of the effects of regions scattered throughout the genome on the trait. However, the underlying biology in these latter cases often remains elusive (Gondro, van der Werf, & Hayes, 2013).

The GWAS have three stages: quality control (QC), association analysis and gene annotation. QC refers to the preparation and filtering phase of the information, where it is sought to eliminate data that generate spurious associations. Poor quality information includes markers and samples with low genotype accuracy. It also comprises SNPs with a low frequency of the minor allele, which is the allele of the marker that has the lowest frequency in the population. The low-quality information includes markers in Hardy-Weinberg imbalance. Quality control also consist of removing SNPs associated with the genetic origin of the individuals, which can create a population stratification (Gondro, van der Werf, et al., 2013).

Most common association analysis models between SNPs and phenotypes depends on statistic tests applied to each SNP. The statistical test to use depends on the trait and productive environment. This test can be chi-square, Bayesian models, multivariate variance analysis (MANOVA), maximum likelihood estimation, Fisher's exact test, or models which include regressions (Ball, 2013; Fernando & Garrick, 2013; Hayes, 2013; Purcell et al., 2007b). Some models adjust the effect of the association between SNPs and

genetic origin (Wang et al., 2014). A multiple test correction is applied to the results of the association to keep false positives below a certain threshold. Among the most used multiple test corrections are Bonferroni and the Benjamini-Hochberg (also called "false discovery rate" (FDR) (H. Zhang et al., 2012). Finally, the annotation refers to the biological contextualization of the genes close to the SNPs associated in the previous step (Thomas, 2017). Tools that integrate quality control, association analysis, and gene annotation are required.

The literature does not report a methodology that integrates the three mentioned steps using the programming language R. Wrong decisions in any of these steps can lead to incorrect conclusions. However, no global method includes all the steps and leads to interpretable and understandable results by users, who do not always have in-depth knowledge in bioinformatics or biostatistics. These stages can be integrated using the environment and language R, which, because it is open-source and robust for statistical analysis, can be used worldwide.

The present work standardized a methodology for quality control, GWAS and gene annotation in highly polygenic traits in ruminants using the environment and language R.

# 1. Objectives

## 1.1 General objective

To standardize a methodology of association of single nucleotide polymorphisms with phenotypes in ruminants and the subsequent biological contextualization of the related genomic regions.

## 1.2 Specific objectives

- To compile and apply tools to filter SNPs and individuals according to the quality of the information they contribute to the association study and apply them within the target populations.
- To detect associations between SNPs and phenotypes.
- To contextualize biologically the regions of the genome associated with the phenotypes.
- To integrate all the analyzes in a workflow.

# 2. Background

## 2.1 Single nucleotide polymorphism (SNP)

The genome-wide association studies (GWAS) have the aim of finding, between thousands of genomic markers, those associated with a phenotype of interest. Typically, only a few markers have large enough effects on the phenotype (Bouwman et al., 2012).

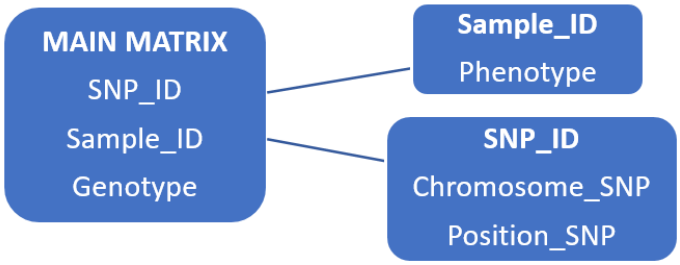
GWAS take advantage of the linkage disequilibrium (LD) between genetic markers and the genes associated with a phenotype. LD is a measure of the association of alleles on gametes or chromosomes (Hudson, 2004). A population in linkage equilibrium in a loci group has their alleles independently distributed in the chromosomes (Hudson, 2004). The LD is stronger between closer markers because they have the same or similar ancestral genealogies and this situation induces a greater dependence between alleles of different markers (Pritchard & Przeworski, 2001). GWAS exploits the LD between markers and gene mutations that determine the phenotypic variations. These mutations are in quantitative trait loci (QTL) when the phenotype is a quantitative trait. These associations arise because there are small segments of the chromosome in the current population that descend from the same common ancestor. The chromosome segments, which come from the same common ancestor without the intervention of recombination, will carry identical alleles of markers or marker haplotypes. If there is a QTL somewhere within the segment of the chromosome, they will also carry identical QTL alleles (Hayes, 2013).

The most commonly used genetic markers in GWAS are single nucleotide polymorphisms (SNPs). These markers correspond to the variation in the sequence in a single nucleotide within a DNA sequence and are commonly the result of transition type mutations (A for G, T for C), although there are also transversions (G or A for T or C) ) and deletions of a single base (Zaid, Hughes, Porceddu, & Nicholas, 2001).

## 2.2 Preprocessing and quality control

Within the preprocessing, the researchers give to the data the format required by the software used in the following steps. Figure 2-1 describes the essential information necessary to conduct a quality control (QC) and GWAS. Additionally, pedigree, environmental covariates, and genotyping quality information are also usually included. The essential steps in QC are analyses of SNP, samples, and batches. SNP quality control includes the genotype quality, and the analyses of population stratification, Hardy-Weinberg equilibrium (HWE), and minor allele frequency (MAF). Sample quality depends on genotype quality, B Allele Frequency (BAF) variance analysis, and missingness and heterozygosity. The batch quality control focuses on the analysis of missing call rate (MCR).

**Figure 2-1:** Essential information necessary to conduct a GWAS.




The genotype file's format and information vary among the companies that carry out the genotyping. In the Illumina's BeadArray technology, each SNP call has a genotype quality score, the GenCall score (Oliphant, Barker, Stuelpnagel, & Chee, 2002). GenCall score correlates with the accuracy of the genotyping call, which Illumina evaluated regarding concordance, reproducibility, and strand correlation (Oliphant et al., 2002). With this technology, it is also possible to generate the quality data call B allele frequency (BAF) and Log R Ratio (LRR) for each sample and each SNP. BAF, cannot be confused with MAF. BAF is the frequency of the B allele in the population of cells of the extracted DNA. The frequency of allele B is expected to be 0, 0.5 or 1. However, the observed frequencies can vary in cases of allelic imbalances, as is the case of trisomic cell populations, where the frequencies can be 0, 0.33, 0.67 or 1. Furthermore, variations in LRR through a

chromosome indicate possible duplications or deletions. It is an indicator of the overall quality of the sample (Gogarten et al., 2012).

The GeneSeek company (https://genomics.neogen.com) uses the Illumina's BeadArray technology. Table 2-1 describes the results this company delivers. These files are compressed twice, first each file and then all these together.

**Table 2-1:** Files structure delivered by GeneSeek.

| File | Column name: Description |
|---|---|
| LocusSummary.csv | Locus_Name: Id or SNP name<br>Illumicode_Name: Illumina code<br>AA_T_Mean: Normalized theta angles mean for the AA genotype.<br>AB_T_Mean: Normalized theta angles mean for the AB genotype.<br>BB_T_Mean: Normalized theta angles mean for the BB genotype.<br>AA_R_Mean: Normalized R value mean for the AA genotypes.<br>AB_R_Mean: Normalized R value mean for the AB genotypes.<br>BB_R_Mean: Normalized R value mean for the BB genotypes. |
| FinalReportCNV.csv | SNP Name: Id or SNP name<br>Sample Id<br>B Allele frequency (BAF).<br>Log R Ratio (LRR). |
| FinalReport.txt | SNP Name: Id or SNP name<br>Sample Id<br>Allele1 – AB: Allele 1 with nomenclature AB<br>Allele2 – AB: Allele 2 with nomenclature AB<br>GC Score<br>X: X intensity<br>Y: Y intensity |
| SNP_Map.txt. | Index: consecutive number<br>Name: Id or SNP name<br>Chromosome<br>Position: chromosome position<br>SNP: alleles with nomenclature A, C, T, G. [Allele 1/Allele 2]. |

When working with large populations, even small sources of systematic or random error can cause spurious associations. QC is a crucial step to avoid erroneous results in the association analyses between SNPs and phenotypes (Wiggans et al., 2009). Therefore, QC must be carried out (C. Laurie, Doheny, Mirel, & Pugh, 2010). In addition to QC based

on genotype quality scores provided by the genotyping company, it is useful to analyze the population stratification, the HWE, and MAF for each SNP, the missingness and heterozygosity of the samples, and the missing call rate (MCR) per batch. Population stratification is aboard on section 2.2.1. Under the assumptions of Hardy-Weinberg equilibrium, it is possible to know the allelic frequencies in the next generation of animals. If the observed values vary significantly from those expected, it is said that the marker is not in Hardy-Weinberg equilibrium (Turner et al., 2011). Hardy-Weinberg imbalance markers may indicate genotyping errors, population stratification and even association with the phenotype (Turner et al., 2011). With the HWE values per SNP, it can be calculated the inbreeding coefficient per SNP, which can show a possible population substructure. A distribution of inbreeding coefficients centered around 0 indicates there is most likely no significant population substructure (Gogarten et al., 2012). MAF is the frequency of the allele less common of a marker across all the population. The exclusion of SNPs with low MAF, generally between 1 and 5%, avoids the association between SNPs and phenotypes without strong statistical support (Gondro, Lee, Lee, & Porto-Neto, 2013).

Missingness per sample is the proportion of missing SNPs in each chromosome of each sample, and its analysis allows to identify and remove samples with significant missing markers (Gogarten et al., 2012). Heterozygosity per sample is the proportion of heterozygous SNPs (Gondro, Lee, et al., 2013). Samples with high heterozygosity could indicate that they are mixed samples (Gogarten et al., 2012). MCR is either the fraction of missing calls per SNP over samples or the fraction per sample over SNPs, and it can be used as an indicator of the batches' genotype quality (C. C. Laurie et al., 2010).

A pipeline was developed to control the quality of data coming from the SNPs of Illumina, which includes population parameters such as MAF and HWE (Gondro, Porto-Neto, & Lee, 2014). A tool commonly used to conduct cleaning of human SNP data is the PLINK software (Purcell et al., 2007a). Marras et al. (2017) developed *Zanardi*, a tool for Linux and Mac environments to integrate files with different formats with the aim of being used in the genomic analysis. This tool works in diploid species that have less than 60 chromosomes. No specific tool for pre-processing non-human species data that performs all the analyses previously mentioned exist. In practice, many GWAS carried out in zootechnical species

make quality control excluding animals that have a percentage of the missing genotype that the authors consider high. They also select SNPs that have a call rate of more than 90% or 95%, an arbitrary p-value of the chi-squared test for the HWE and a MAF higher than a value that each study considers adequate (Eusebi et al., 2017; Fortes et al., 2010; Nishimura et al., 2012; C. Zhang et al., 2015).

### 2.2.1 Population stratification

The importance of determining population stratification lies in avoiding associations between SNPs with phenotypes, when in fact they are associated with the genetic origin of individuals. Therefore, not taking into account the structure of the population can cause false positives (Pritchard et al., 2000). In important zootechnical species, the GWAS should consider that individuals are generally highly related due to artificial selection, and therefore there is a tendency for a population stratification higher than this in human populations. In sheep and cattle, the selection of the parents and the direction of the crosses made by the human favors the presence of spurious associations. In this sense, if for example a parent has a desirable phenotype and it is also homozygous for a rare allele not associated with the phenotype, their offspring may present the desirable phenotype and a high frequency of the rare allele. In this way, the researchers could mistakenly associate the rare allele with the desirable phenotype.

Different methodologies have been developed to avoid false positives generated by population stratification. Unlinked genetic markers have been used to infer details of the population structure and estimate the ancestry of the individuals sampled, and then that information is used to classify individuals into subpopulations and make association tests between them (Pritchard et al., 2000). One of the methodologies that have been used to determine ancestry and classify individuals into subpopulations is cluster analysis with multivariate techniques, including multiple correspondence analysis (MCA) (Cifuentes, Cortés, Franco, & Niño, n.d.). In this case, an MCA is conducted between the SNPs and the genetic origin, and the result is plotted. If the clusters of genetic origins overlap, there are no populations stratification (Cifuentes et al., n.d.). Another widely used methodology is the transmission disequilibrium test (TDT), which depends on genotyping not only individuals who present the phenotype of interest but also their parents and thus divide individuals by families. TDT identifies the effect of the alleles of each marker that are in

each family on the phenotype (Spielman, McGinnis, & Ewens, 1993). In important zootechnical species, implementing the TDT methodology is unfeasible due to the cost of genotyping all the parents (Hayes, 2013). Principal component analyses also are used to avoid population stratification (Y. Zhang, Guan, & Pan, 2013).

Mixed model algorithms, instead of dividing the population into subpopulations or families, include all the individuals in the association and treat the genetic origin as a random effect of the model (Y. S. Aulchenko, de Koning, & Haley, 2007; Hayes, 2013; H. M. Kang et al., 2008; Hyun Min Kang et al., 2010; Lippert et al., 2011; Listgarten, Lippert, & Heckerman, 2013; Svishcheva, Axenovich, Belonogova, van Duijn, & Aulchenko, 2012). Some researchers implement these models in the PLINK software (Rentería, Cortes, & Medland, 2013). Additionally, multiple component analysis to identify clusters of genetically related individuals have been developed (Jombart, Devillard, & Balloux, 2010).

## 2.3 Association analysis

The association with categorical phenotypes is done applying a chi-squared test for each SNP concerning the phenotypic state (Purcell et al., 2007a). Logistic regressions are also used to conduct the association with categorical phenotypes, and linear regressions are used when the phenotypes are continuous (Hayes, 2013). The Wald test is applied to the result of the regressions to evaluate the degree of contribution of each SNP to the phenotype.

The chi-square test ($X^2$) allows to determine the association between two variables by comparing the expected and observed frequencies. The null hypothesis ($H_0$) is "there is no association between the variables" and the alternative hypothesis ($H_1$) is "there is an association between the variables" (Pita & Pértega, n.d.). In general, for a table of $i = 1, \dots, r$ rows per j $= 1, \dots, k$ columns, the value of the statistic $X^2$ is calculated as follows:

$$X^2 = \sum_{i-1}^{r} \sum_{j-1}^{k} \left(O_{ij} - E_{ij}\right)^2 \Big/ E_{ij}$$

Where:

$O_{ij}$ denotes the observed frequencies. It is the number of observed cases classified in the $i^{th}$ row of the $j^{th}$ column.

$E_{ij}$ denotes the expected or theoretical frequencies. It is the number of expected cases corresponding to each row and column. It is the observed frequency when the two variables were independent. The expected values are the product of the marginal totals divided by the total number of cases $(n)$.

Subsequently, one compares the $X^2$ value obtained with the value of the statistical distribution. The latter depends on a given alpha value and the degrees of freedom $(r-1)(k-1)$. Then, if the calculated value is greater than the value of the statistical distribution, $H_0$ is rejected.

For the associations made with regressions, the most straightforward methodologies are the single marker regression models, in which a randomly mating population without population stratification is assumed, and has the following model according to Hayes (2013):

$$y = Wb + Xg + e$$

Where $y$ is the phenotype vector, $W$ is a design matrix that assigns registers to fixed effects of the phenotypes, $b$ is a vector of fixed effects, $X$ is a design matrix that assigns registers to the effect of the marker, $g$ is the effect of the marker, and $e$ is a vector of random deviations $e_{ij} \sim N(0, \sigma_e^2)$, where $\sigma_e^2$ is the variance of the error. In this model, the marker is a fixed effect, and the model is additive since two copies of the rare allele have twice the effect of a copy and a genotype without copies of the rare allele does not have any effect.

After performing the regression, it is necessary to test whether $g$ is zero or not. The null hypothesis is that the marker is not associated with the phenotype. The Fisher's exact test or one of the three asymptotically equivalent tests, i.e., likelihood ratio test, score test, and Wald test, can be applicable (Dominik, 2013; Hayes, 2013). Wald test is considered the gold standard test (Yu, Demetriou, & Gillen, 2015). Given that the parameter of interest is zero, the Wald test is reduced from the general expression explained by Wasserman (2004). Then, the Wald test used is the quotient between the marker's estimated coefficient

$(\hat{g})$, and its standard error $se(\hat{g})$. This statistic follows a chi-square distribution with one degree of freedom. The R package GWASTools uses the Wald test to evaluate the significance of associations between SNPs and phenotypes (Gogarten et al., 2012). The function of the Wald test is following described.

$$Wald = \hat{g} / se(\hat{g})$$

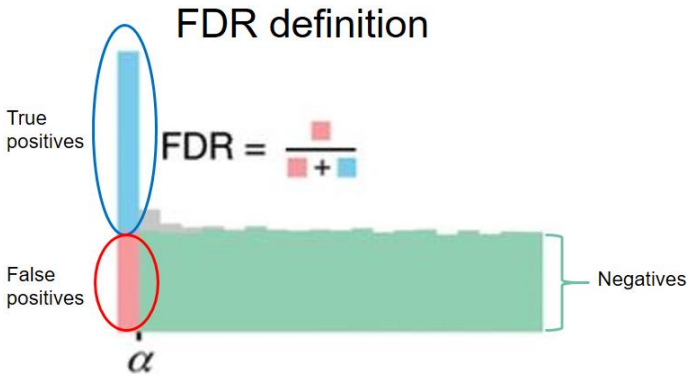
In addition to the marker-by-marker models, algorithms to make associations using linear models, eliminating the population structure have been developed. The two-step algorithm genomewide rapid association using mixed model and regression (GRAMMAR), implemented in the R package of GenABEL, uses a linear mixed model to make the association between genetic markers and the phenotype adjusted by the effects of the family (Y. S. Aulchenko et al., 2007). The variation of GRAMMAR, GRAMMAR-Gamma has been widely used in ruminants(Svishcheva et al., 2012). EMMA is an algorithm implemented in the R language that makes the association between SNPs and phenotypes corrected by the population stratification and the relationship between individuals (H. M. Kang et al., 2008). However, EMMA is computationally infeasible to do association analysis with thousands of markers because it calculates the variance parameters for each marker. The EMMAX algorithm was developed to solve this problem. EMMAX assumes that since the effect of each marker is very small, the variance parameters can be estimated only once for the entire database (Hyun Min Kang et al., 2010). Subsequently, another EMMA variant was developed, called GEMMA, which is also faster than EMMAX and allows analysis of association with thousands of markers (X. Zhou & Stephens, 2012). In addition to the linear models, Bayesian models have also been proposed to make the association (Grimm, 2015).

## 2.4 Multiple comparisons

The number of multiple tests presents a challenge when one defines the level of significance (α) of the associations between SNPs and phenotypes. For instance, by defining a "nominal" α of 5% for a 50,000 SNP chip, 2,500 false positives could be present by chance (type I error). Multiple testing corrections reduce the presence of these errors

modifying the p-values of the associations. Bonferroni method divides the type I error by the total number of tests, and it is considered a conservative test (Bonferroni, 1936). Another correction widely used is the Benjamini-Hochberg method that seeks to control the error rate called "false discovery rate" (FDR) (Benjamini & Hochberg, 1995). First, it separates the false positives from the true positives. To do so, it orders all positives (true and false) from lowest to highest according to their p-value, generates a ranking with this information and separates the p-values into two groups according to their position in the ranking. The group of true positives is made up of the smallest p-values, counted from the lowest (number 1 in the ranking) to this one that occupies the position equal to the number of expected false positives. The group of false positives consists of the highest p-values, counted from the number of predicted false positives. The logic behind this differentiation between true and false positives is the p-values distribution (Figure 2-2).

**Figure 2-2:** Definition of FDR. Adapted from Krzywinski y Altman (2014).




After separating the true from the false positives, the methodology groups false positives and negatives. Then, the method orders the values upwards, generates a ranking of them, and assigns a p-value of their position in the ranking. Then the next function is applied to the larger p-value.

$$Q_e = E\left(\frac{V}{R}\right),$$

where

$E$ is the p-value before the adjustment,

$V$ is the range or number of elements in the group. It is considered zero when R = 0.
$R$ is the position of the p-value in the ranking.

Since for the larger p-value $V$ and $R$ are equal, their $Q_e$ will be the same, and this one will be the adjusted p-value. Then the same function is applied to the second largest p-value, and the smallest number between its $Q_e$ and the adjusted p-value of its predecessor is the adjusted p-value, that in this case is the adjusted p-value of the greater p-value. The methodology applies this same calculation to all the p-values in descendent order (Starmer, 2017).

## 2.5 Manhattan plot

The Manhattan plot is a typical tool to visualize the association result. In this plot, the chromosomes and the position of the marker within them are on the x-axis, and the y-axis has the negative Log base 10 of the p-value of the association between the marker and the phenotype (Figure 2-3). A SNP is considered associated with the phenotype if its value is higher than a given cutoff, which is generally between 0.4 and 0.7. The plot shows this cutoffis as a dotted line. Another plots are the “Conditional Manhattan plots”, which uses the p-values after multiple comparison adjust (Figure 2-3). In these plots, for instance, the SNPs with conditional –log10 FDR greater than 1.3 (that is FDR<0.05), are considered as associated with the trait (Andreassen et al., 2013).

**Figure 2-3:** Example of a Manhattan plot. Taken from (Ren et al., 2016).

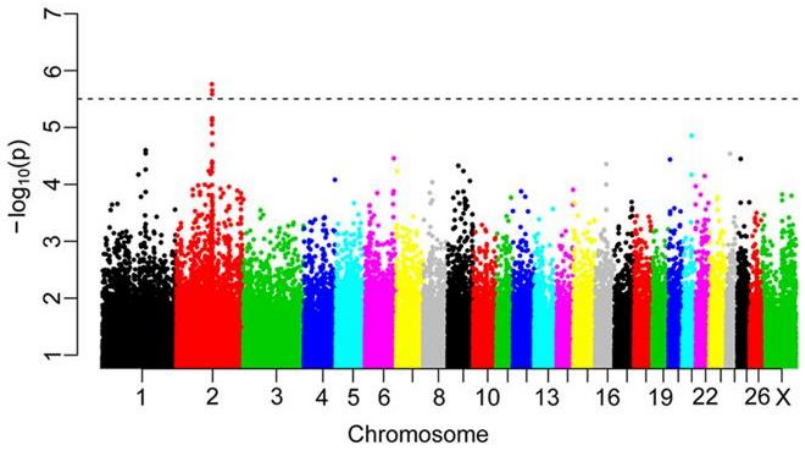
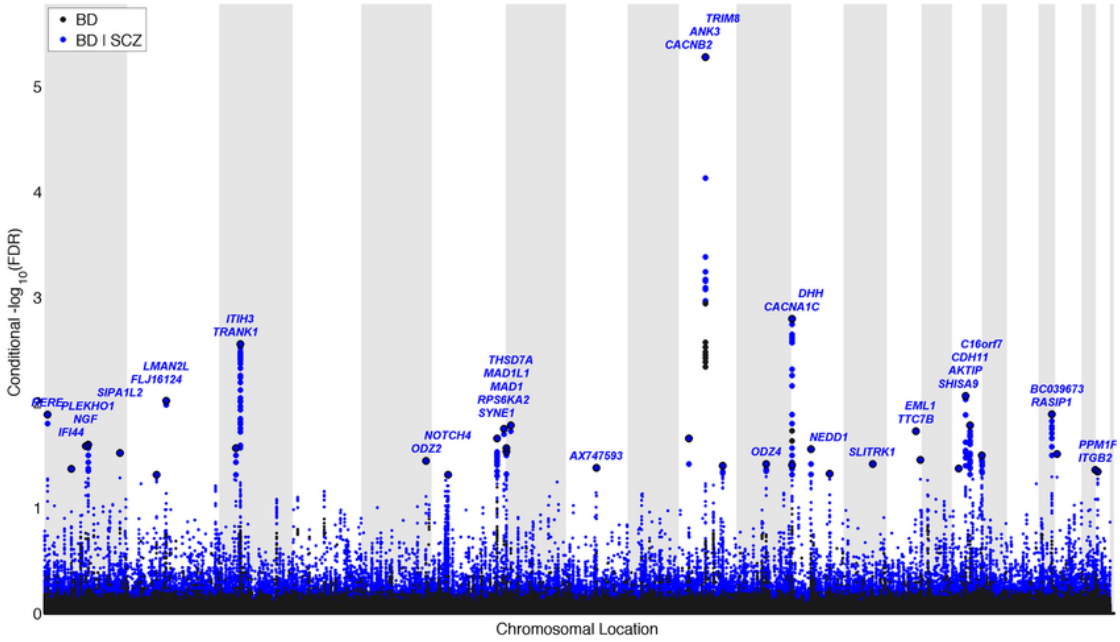


**Figure 2-4:** “Conditional Manhattan plot” of conditional −log10 (FDR) values. (Andreassen et al., 2013).




## 2.6 Gene annotation

The aim is to know the biological function of the genes possibly associated with the phenotype of interest. The SNPs identified as potentially associated with the phenotype in the GWAS can be in a gene or an intergenic region. If the marker is in an intergenic region, it is likely to be in linkage disequilibrium with nearby genes. Open access databases contains biological function information about genes. Within the primary databases are those administered by the National Center for Biotechnology Information (NCBI), the European Institute of Bioinformatics (EMBL-EBI) and the UniProt consortium, as well as the Encyclopedia of Genes and Genomes of Kyoto (KEGG), and the Gene Ontology (GO) project. It is possible to access the information found in these databases directly on the websites, as well as by other ways, as R packages.

The biomaRt package accesses the Ensembl, COSMIC, Uniprot, HGNC, Gramene and Wormbase databases (Durinck et al., 2005; Durinck, Spellman, Birney, & Huber, 2009). The Category package allows performing analysis of GO terms (Gentleman, 2018b). KEGGrest allows accessing to the metabolic pathways in which the proteins encoded by the genes of interest are involved (Tenenbaum, 2018). The annotate library accesses the

NCBI databases (Gentleman, 2018a). mygene accesses all the bases mentioned and organizes the result according to the identifiers of the genes, transcripts, proteins and GO terms used in each one of them (Mark, Thompson, Afrasiabi, & Wu, 2018).

# 3. Proposed methodology

## 3.1 Introduction

Genome-wide association studies (GWAS) look for the association between genomic markers and phenotypes. Before performing these analyses, it is necessary to do file input and quality control (QC) of data. Additionally, it is essential to assess the underlying biology of the associations between markers and phenotype. Commonly, the researchers annotate the genes close to the associated markers. It is possible to use R libraries developed to study human data. The present work proposes a workflow to performs all these steps.

Before performing a GWAS, it is necessary to preprocess the genotype files and do quality control (CQ) of data. Generally, genotypes files are large, and genotype companies deliver them with specific formats. These files demand an efficient way to input their information into the software that will be used in the analyses. After input data, QC filters out markers and individuals that can conduct to spurious associations. Finally, is conducted the association and multiple comparison analyses, as well as the gene annotation.

It is possible to adapt tools designed to perform QC, GWAS, and annotation for human data, to do the same in non-human diploid species. There are many tools to perform these analyses in human data. Adapting these tools can be useful in disciplines like agriculture and biology. Not everyone interested in GWAS results has the background and the experience to carry out these analyses. A portion of this workflow is the adaptation of the R package GWASTools to perform analyses of non-human data.

The present work created the Diploid-GWAS tool. Diplod-GWAS has two modules for conduct GWAS of any diploid species using the R environment (R-Team, 2013). Module one is for input, QC and association analysis. Module two is for gene annotation. Furthermore, we submitted to publication to the journal Animal Genetics a paper with the

workflow and the repository. The work used two sets of toy data to test the code, one within the GeneSeek structure, and other with the general structure.

The toy data and code are in the Appendix A, as well as freely available in the repository https://github.com/bojusemo/Diploid-GWAS. This material is divided into three sections: Module one QC and association - general structure, module one QC and association - GeneSeek structure, and Module two - annotation. The section *Input and files structure* describes the data files of the module one.

## 3.2 Methodologic article submitted to the journal Animal Genetics

**Title:** Original: Diploid-GWAS: An R workflow for quality control, GWAS, and annotation in diploid species

**Authors**: Boris Sepúlveda-Molina, Liliana López-Kleine

**Summary:** There are less free tools for genome-wide association studies in non-human organisms than in *Homo sapiens.* Here, we present a workflow that adapts and integrates currently available R tools into a workflow to perform quality control and filtering of SNP data of all diploid organisms, as well as the association between phenotype and SNPs, and, finally, the annotation of the genes close to associated markers. The code and toy data are freely available.

**Keywords:** chip; genome-wide association studies; single nucleotide polymorphism, annotation.

**Description:** There are many analyses associated with genome-wide association studies (GWAS). GWAS allow associating single nucleotide polymorphism (SNP) with phenotypes of interest (Gondro, van der Werf, et al., 2013). Before conducting a GWAS, it is necessary to implement a quality control (QC) on the data, which includes removing genotypes with low accuracy, detecting and correcting population stratification, and performing Minor Allele Frequency and Hardy-Weinberg equilibrium testing of markers. Then, a statistical test to associate SNPs and traits of interest, followed by an adjustment of the statistical significance of these associations using a multiple testing correction is applied. Several association tests exist, such as chi-square, Bayesian models, analysis of variance, Fisher's exact test, and significance tests on regression models (H. Zhang et al., 2012). Most common multiple testing corrections are Bonferroni and false discovery rate (FDR) (H. Zhang et al., 2012). After the GWAS, and to study the underlying biology of phenotype-genotype association, is conducted a gene annotation analysis of the genes close to the associated markers.

In the R environment (R-Team, 2013) are available tools to perform QC, GWAS, and annotation, but none integrate these three steps into a workflow. Moreover, most existing libraries in R and other available software perform GWAS in humans. The R package GWASTools, for example, executes QC and association analysis from human data (Gogarten et al., 2012). There are also packages directed to analyze no human data. Gondro et al. (Gondro et al., 2014) developed a pipeline for QC of Illumina genotypes. The package GenABEL performs QC and GWAS (Yurii S. Aulchenko, Ripke, Isaacs, & van Duijn, 2007). GenABEL's QC is related to genotypes that pass a given call rate, redundancy, minimal marker allele frequency and deviation from Hardy–Weinberg equilibrium (Yurii S. Aulchenko et al., 2007). GenABEL and GWASTools support their GWAS analysis in the standard R procedure of generalized linear models (glm). GWASTools also uses the Wald test for determining the significance of regression coefficients. R core team produces functions to adjust P-values for multiple comparisons, including Bonferroni and FDR. Additinally, there are packages for gene annotation, as is the case of the package mygene (Mark et al., 2018).

The workflow presented in this paper is available in https://github.com/bojusemo/Diploid-GWAS and performs QC, association analysis, and gene annotation (Figure 3-1). The process undertakes the following steps:

1. The user provides the paths of the GeneSeek genotypes that use Illumina technology, and the Scan annotation data frame. The last file must include the subjectID, sex, and phenotypic traits.

2. Genotype information is extracted from GeneSeek files. GeneSeek delivers a compressed .zip main folder, within which are files compressed again in .zip. The main files extracted are SNP_Map.txt, Sample_Map.txt, LocusXDNA.csv, LocusSummary.csv., FinalReport.txt, and DNAReport.csv. Each file is preceding by the structure "*_**_" where "*" correspond to the name of the customer and "**" is the delivery date in format YYYYMMDD (Y = year, M = month, D = day).

3. Input the files with genotypes and scan annotation into R. In the genotype files, the user has the option of select the number of rows to remove from the header.

4. Reorder genotype information. QC is performed with the package GWASTools (Gogarten et al., 2012). While GeneSeek files include data from many samples, GWASTools requires each sample's genotype in a single file. For that reason, a filer per sample is performed and exported.

5. GWASTools objects Creation. The code creates the metadata objects "SNP Annotation Data Object" and "Scan Annotation Data Object." The SNP Annotation Data Object stores information of each SNP: unique integer ID; Illumina name; the chromosome, which can be the number of an autosome or adding 1, 2, 3, or 4 to the last autosome number if the chromosome is X, Y, mitochondrial, or unknown, respectively; the base pair position on chromosome; the allele in A/B format; and the quality information "mean theta for AA cluster", "mean theta for AB cluster", "mean theta for BB cluster", "mean R for AA cluster", "mean R for AB cluster", "mean R for BB cluster". The Scan Annotation Data Object stores: sample ID also known as scan ID, subject ID, sex, genetic origin, and phenotypes.

6. Export metadata with the format of *genomic data structure* (GDS). GDS format allows for efficient memory management for GWAS (Zheng et al., 2012). The workflow creates three files that associate the information of GWASTools objects with genotype information of each SNP of each sample. The first one stores the alleles; the other two have quality variables: call rate, x and y position, BAlleleFreq, LogRRatio.

7. QC process. Reports in PDF and text files about the quality per batch, sample, and SNP are generated.

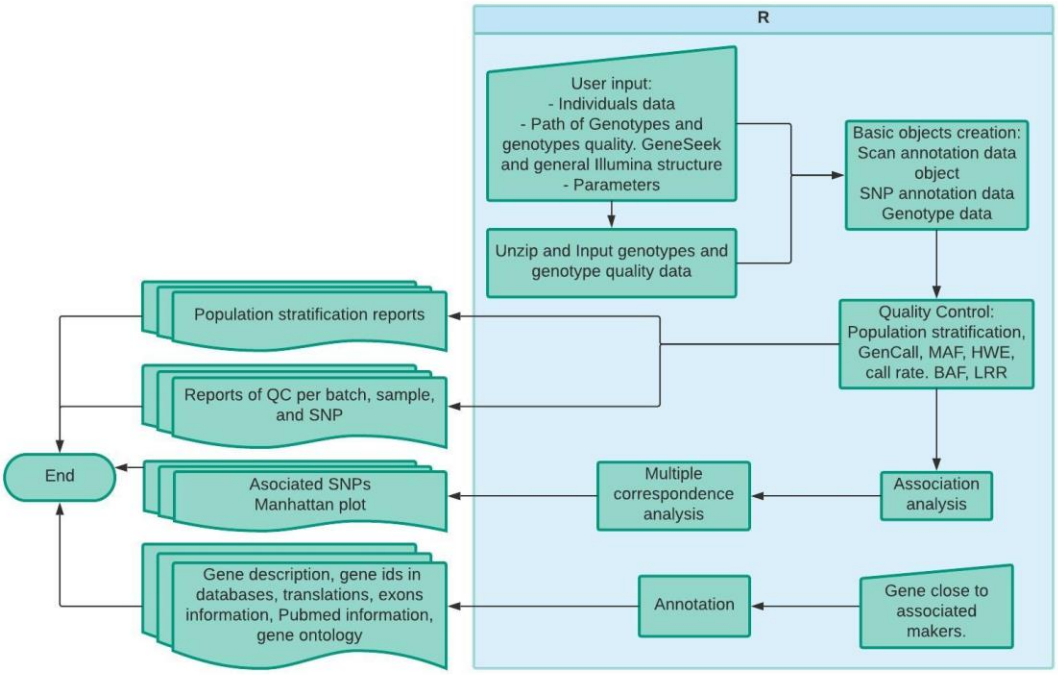
8. Population stratification analysis. The workflow uses multiple correspondence analysis to detect population stratification using the function `dudi.acm` from package ade4 (Dray & Dufour, 2007a). The code plots the result with the function `fviz_mca_ind` of the package factoextra (Kassambara & Mundt, 2017). If

population stratification is detected, the code performs an association analysis with logistic regression between genetic origin and SNPs. Then, the workflow removes the associated SNPs. Then, the code makes a new multiple correspondence analysis with the retained SNPs and plots the result.

9. Association analysis. The tool carries out the association analysis with GWASTools with logistic and linear regressions for categorical and continuous phenotypes, respectively. The code determined the significance of the regression coefficient with the Wald test. The code adjusts the p-values with the multiple correspondence analysis tests Bonferroni or FDR with the function `p.adjust` of the package stats (R_Core_Team, 2018). This test reduces the presence of type I error. The workflow exports the associated SNPs and the Manhattan plot created with the function `assocRegression` of GWASTools.

10. Annotation analysis. The user inputs the Entrez or Ensembl gene ID that want to analyze. The code performs the gene annotation with the function `getGene` of the package mygene and exports the information generated with this package (Mark et al., 2018).

**Figure 3-1:** Workflow.




## 3.3 Description of module one

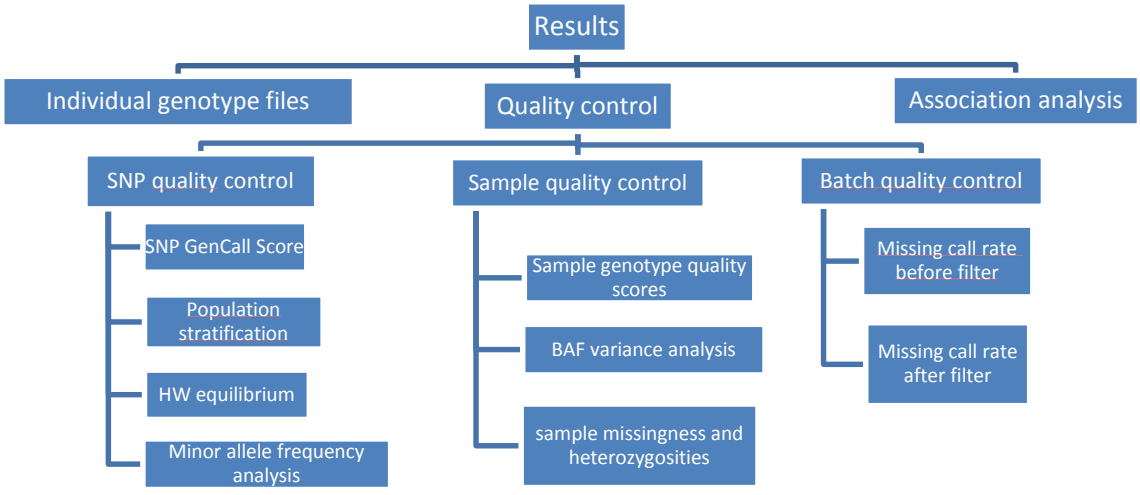
Module one performs data input, QC and association analysis. There are two versions of this module, one for files with GeneSeek structure, and another for files with a general structure. The code differences regard the way to input the data, the necessary object creation and, the analysis of the population stratification. The remaining code is the same for both versions. The R packages used are GWASTools (Gogarten et al., 2012), ade4 (Dray & Dufour, 2007a), and SNPRelate (Zheng et al., 2012). The examples of the results show in this module were obtained applying the code in the toy data described in the Apendix A.

We developed this tool for users that do not necessarily have worked with R/Bioconductor. The user must install R with the instructions of https://cran.r-project.org/. Then, they can install a graphical interface (www.rstudio.com). Files are created and stored in the same environment as the script. Therefore, it is recommended to store the script in a new folder. The user must install Bioconductor and some packages. Bioconductor can be installed typing in the console of RStudio the command

`source("https://bioconductor.org/biocLite.R")`. Then, packages must be installed typing in the console the following commands: `biocLite("GWASTools")`, `biocLite("SNPRelate")`, `install.packages("ade4")`, `install.packages("factoextra")`, and `install.packages("mygene")`. The user must define the parameters as described in this document in the section *Parameter input*. The user also must input the data as described in the *Input and files structure* section of this document. Then, select all the script (Ctrl + A) and run (Ctrl + Enter or click on *run*). The results will appear in the directory *Results*, created in the same path where is executed the script.

The workflow carries out three kinds of results related to quality control (QC): SNP quality control, sample quality control, and batch quality control. Low-quality data are filtered out. A batch is a group of samples genotyped at the same time. The batch quality control allows the researchers to decide whether to remove batches. The folders described in Figure 3-2 store the results.

**Figure 3-2:** Results folder structure.



### 3.3.1 Input and files structure

**Genotypes with the GeneSeek structure**

GeneSeek works with Illumina technology and delivers the genotypes with the structure following described. These files are compressed. The user must specify in quotation marks, the genotype´s path in the object `path.folders`. For instance, if genotypes are in a folder called *Data*, which is in the desktop, the path would be `path.folders <- "~/Desktop/Data"`.

These files must have at least the columns described below and in the same order. They can contain more columns and different numbers of header rows.

- LocusSummary.csv
  - Locus_Name: Id or name of the SNP
  - Illumicode_Name: Illumina code name
  - AA_T_Mean: Normalized theta angles mean for the AA genotype.
  - AB_T_Mean: Normalized theta angles mean for the AB genotype.
  - BB_T_Mean: Normalized theta angles mean for the BB genotype.
  - AA_R_Mean: Normalized R-value mean for the AA genotypes.
  - AB_R_Mean: Normalized R-value mean for the AB genotypes.
  - BB_R_Mean: Normalized R-value mean for the BB genotypes.

- FinalReportCNV.csv. This file is optional.
  - SNP Name: Id or name of the SNP
  - Sample ID
  - B Allele Freq: Allelic intensity ratio. B Allele frequency (BAF).
  - Log R Ratio: Genotyping total signal intensity. Log R Ratio (LRR).

- FinalReport.txt
  - SNP Name: Id or name of the SNP
  - Sample ID
  - Allele1 – AB: Allele 1 with nomenclature AB
  - Allele2 – AB: Allele 2 with nomenclature AB
  - GC Score: GeneCall score. It is a measure of the genotyping quality.

    - X: intensity value X
    - Y: intensity value Y

- SNP_Map.txt. The third column must be the Chromosome.
    - Index: consecutive number
    - Name: Id or name of the SNP
    - Chromosome: SNP chromosome
    - Position: SNP position on the chromosome
    - SNP: alleles with nomenclature A, C, T, or G. [Allele 1/Allele 2].
    - ILMN Strand: Order of Alleles A and B
    - Customer Strand: Order of Alleles A and B

After inputting genotypes, the user must input a file with sample information concerning to phenotypes and population structure. The R object created must be called `Scan_Annotation_Data_Frame`. The file must have these columns in this order:

- subjectID: It must coincide with the column *Sample ID* of the file *FinalReport.txt*.
- sex: coded as M for male and F for female.
- trait: phenotype

Additionally, it could have the following columns:

- batch: genotyping batch
- genetic_origin: genetic origin group coded as 0 or 1 for each origin.
- sire: Male parent
- dam: Female parent
- generation: the number of the generation of the subject

**Genotypes with the general structure**

This version aims to allow the input of data with a general structure. The user must have data frames and generate the following R objects with them. The *Import Dataset* section of RStudio can be used to do so, activating the option of defining the first row as header. It is essential that the files have the same name, and the structure described below.

- *snp_annot_data_frame*. This file must contain these columns in order:

- snpName: Id or name of the SNPs
- chromosome: SNP chromosome
- position: SNP position in the chromosome

Additionally, the following columns could be present:

- alleleA: A, C, T, or G
- alleleB: A, C, T, or G
- AA_T_Mean: Normalized theta angles mean for the AA genotype.
- AB_T_Mean: Normalized theta angles mean for the AB genotype.
- BB_T_Mean: Normalized theta angles mean for the BB genotype.
- AA_R_Mean: Normalized R value mean for the AA genotypes.
- AB_R_Mean: Normalized R value mean for the AB genotypes.
- BB_R_Mean: Normalized R value mean for the BB genotypes.

- *Scan_Annotation_Data_Frame*. This file must contain these columns in order:
  - subjectID: subject identifier
  - sex: sex coded as M=male and F=female
  - trait: phenotype

Additionally, it can have these columns:

- genetic_origin: genetic origin group
- sire: Male parent
- dam: Female parent
- generation: the number of the generation of the subject
- batch: genotyping batch

- *FinalReport*. This file must contain these columns in the following order:
  - SNP.Name: Id or name of the SNPs. It must match with the names of the column snpName of the file SNP annotation data frame.
    - Sample.ID: subject identifier. It must match with the names of the column subjectID of the *scan annotation data frame* file.
    - Allele1...AB: Allele 1 with nomenclature AB
    - Allele2...AB: Allele 2 with nomenclature AB
    - GC Score: GeneCall score. It is a measure of the genotyping quality.

Additionally, it can have these columns:

- X: intensity value X
- Y: intensity value Y

- *FinalReportCNV*. This file is optional. The columns are:
  - SNP.Name: Id or name of the SNPs. It must match with the names of the column snpName of the file SNP annotation data frame.
  - Sample.ID: subject identifier. It must match with the names of the column subjectID of the *scan annotation data frame* file.
  - B Allele Freq: Allelic intensity ratio. B Allele frequency (BAF).
  - Log R Ratio: Genotyping total signal intensity. Log R Ratio (LRR).

## 3.3.2 Parameter input

The users provide the parameters in the R script and store these parameters in the R objects described below. The objects must have the names described in Table 3-1.

**Table 3-1:** Workflow parameters provided by the user.

| Parameter | Description |
|---|---|
| `Autosomes` | Number of autosomes of the species |
| `num_SNP_auto` | Total number of SNPs in autosomes |
| Quality control parameters | |
| `Maf` | Minor allele frequency cutoff |
| `Pvalue` | P-value cutoff of Hardy-Weinberg equilibrium test. A Fisher's exact test is used to determine the deviance of SNPs from Hardy-Weinberg Equilibrium. |
| `Cutoff` | The cutoff of missing call rate |
| `mean_GC.score_sample`<br>`median_GC.score_sample` | GenCall Score cutoff:<br>GenCall Score mean per sample<br>GenCall Score median per sample |

| | |
|---|---|
| `mean_GC.score_snp`<br>`median_GC.score_snp` | GenCall Score mean per SNP<br>GenCall Score median per SNP |
| `pop_signif` | The cutoff of the population stratification's Manhattan plot. This parameter is optional. The user can conduct the genetic origin analysis when the population has two origins. It should be a column called *genetic_origin* in the *Scan_Annotation_Data_Frame* with values 0 or 1 for each origin. |
| Association analysis parameters | |
| `Outcome` | Column name of the phenotype of interest |
| `model.type` | Model type. Can be linear or logistic. Logistic is for case-control studies with values of 0 and 1 in the phenotype column. |
| `Covar` | Covariates. If there are covariates, replace "NULL" with the name or names of covariates. Covariates names must be columns in the *Scan_Annotation_Data_Frame.* For more than one covariate use the form<br>covar = c(covar_name_1, covar_name_1, ...) |
| `Ivar` | Covariate interaction with genotype. |
| `CI` | Confidence interval. |
| `block.size` | Number of SNPs to read in at once. |
| `method` | Multiple comparisons test. Use fdr or bonferroni. |
| `signif` | Manhattan plot cutoff. |

### 3.3.3 Creation of basic objects

This section decompresses the GeneSeek files. The files, that are in the `path.folders` (see I*nput and files structure* section), are uncompressed and imported in R. Additionally, data from both GeneSeek and general structures are imported in R. Objects of the package GWASTools need to be created to do QC steps and association analysis (Gogarten et al., 2012).

Per each batch, GeneSeek delivers many files. This workflow uses the files LocusSummary, SNP Map, FinalReport, and FinalReportCNV. These files are provided individually compressed in a zip format. Additionally, these zip files are compressed again in another zip by each batch. A list of batch paths is created with the function `list.files` to uncompress the zip batch files. The function `lapply` is used to apply the function `unzip` to the files of this list. The code repeats these two steps with a list of the individual zip files. The tool contains functions to input each kind of file. The workflow creates a list per each kind of files from different batches, using the function `list.files`. These functions are `read_function_LocusSummary`, `read_function_SNP_Map`, `read_function_FinalReport`, and `read_function_FinalReportCNV`. The functions are applied to the list using the function `lapply`. Then, the code unifies the objects in the `FinalReport` and `FinalReportCNV` objects with the function `do.call`. The tool removes the SNPs that are not in all the samples. This last step allows working with individuals genotyped with different SNP chips.

GWASTools package was developed to work with human data (Gogarten et al., 2012). This section of the workflow modifies the objects created in GWASTools to use them in the analyses of non-human diploid species. First, the tool generates the SNP Annotation Data and Scan Annotation Data objects. They store metadata of markers and samples, respectively. Then, the workflow makes the Data Files, which store genotypes and genotype quality data with the *Genomic Data Structure* (GDS) format (Zheng et al., 2017, 2012). This format divides data in arrays and allows to store and access to them. With this information, the workflow makes three data files: Genotype File, Intensity File, and B Allele Frequency and Log R Ratio File. The former contains genotypes and the two-last genotype quality data. Finally, there are combined the Annotation objects and Data Files into one GenotypeData and two IntensityData objects.

The SNP annotation object stores marker's information. If the data has a general structure, the user must input this object. With GeneSeek structure data, the object is created merging the SNP Map and LocusSummary files with the function merge. In the chromosome column, the values of X, XY (pseudoautosomal region), Y, mitochondrial, and unknown are reassigned as the number of autosomes plus one, two, three, four, and five, respectively.

Then, the code creates the SNP annotation data frame. This data frame contains the SNP's name, chromosome, position, allele A, Allele B, Bead Set ID, mean theta for AA cluster, mean theta for AB cluster, mean theta for BB cluster, mean R for AA cluster, mean R for AB cluster, and mean R for BB cluster. The SNPs removed from the FinalReport files are filtered out of the SNP annotation data frame to match the information. With this data frame, the tool performs the SNP Annotation Data Object.

The code recodes the SNPs to create the SNP Annotation Data Object. The SNPs are recoded creating a new column called snpID in the `snp_annot_data_frame`. snpID is the key variable to connect the SNP Annotation Data with the GDS files. This identifier is as a consecutive number generated sorting the `snp_annot_data_frame` object by chromosome and then by position. With this data frame, the SNP Annotation Data Object is created using the function `SnpAnnotationDataFrame`. Metadata is added to describe the columns of the SNP Annotation Data Object with the function `varMetadata`.

The workflow also recodes the samples. The user must input the Scan Annotation Data object that stores the information of the samples. The tool performs an identifier for each sample (`scanID`). This id is the key variable to connect the Scan Annotation Data with the GDS files and corresponds to the row number of each sample in the scan annotation data frame. With this new information, and using the function `ScanAnnotationDataFrame`, the workflow makes the Scan Annotation Data Object.

The workflow makes a file with genotype information for each sample to match with the GWASTools structure. First, it creates a folder inside the data directory to store the files. The `scanID` is an index in the creation of the individual files. The workflow generates a data frame for each sample of the FinalReport and FinalReport CNV objects. These files are merged using the function `func_Merg_FinalReport_FinalReportCNV` and sorted in a list called `FinalReports`. Then, this list is joined with the `scanID` in the `FinalReport_scanID` object using the functions `lapply` and `func_Merg_FinalReport_scanID`. At this point, the function `func_create_file_names` makes the files' names and paths. Then, the tool exports the files using the functions `func_creat_files`, `func_individual_files`, and the `apply` family functions. The folder *Individual genotype files* store the files (Figure 3-2).

After the creation of individual files, the tool makes the GDS files. It creates data frames with key variables of the Scan Annotation and SNP Annotation objects. These data frames are required in the GDS files creation to match them with the key variables of individual files. One object is called `scan.annotation` and stores scanID, scanName, and path files. The other one is the `snp.annotation`, that stores the snpID, snpNAme, chromosome and position. Then, the code produces three GDS files with the use of the function `createDataFile`. These files are diag.geno, diag.qxy, and diag.bl. The first one stores the genotype information. diag.qxy has information about the GC score, and X and Y coordinates. The last GDS file contains the BAlleleFreq and LogRRatio.

Finally, the workflow combines these three GDS files with the Annotation objects using the functions `GenotypeData` and `IntensityData` to create the objects used in the QC and association analyses. These objects are `genoData`, `qxyData`, and `blData`. The tool adjusts the number of chromosomes to the chromosome number of the species.

### 3.3.4 Quality control

In this section, the tool conducts the quality control per SNP, sample, and batche. The SNP quality control reports includes the analyses of GenCall (GC) score, population stratification, Hardy-Weinberg equilibrium, and minor allele frequency. The sample quality control generates reports of the GC score, B Allele Frequency variance, and missingness and heterozygosity within samples. Finally, the batch quality control analyzes the missing call rate (MCR). The tool exports the results as text or PDF files, creating the paths with the function `dir.create`.

**SNP quality control**

- GC score analysis per SNP

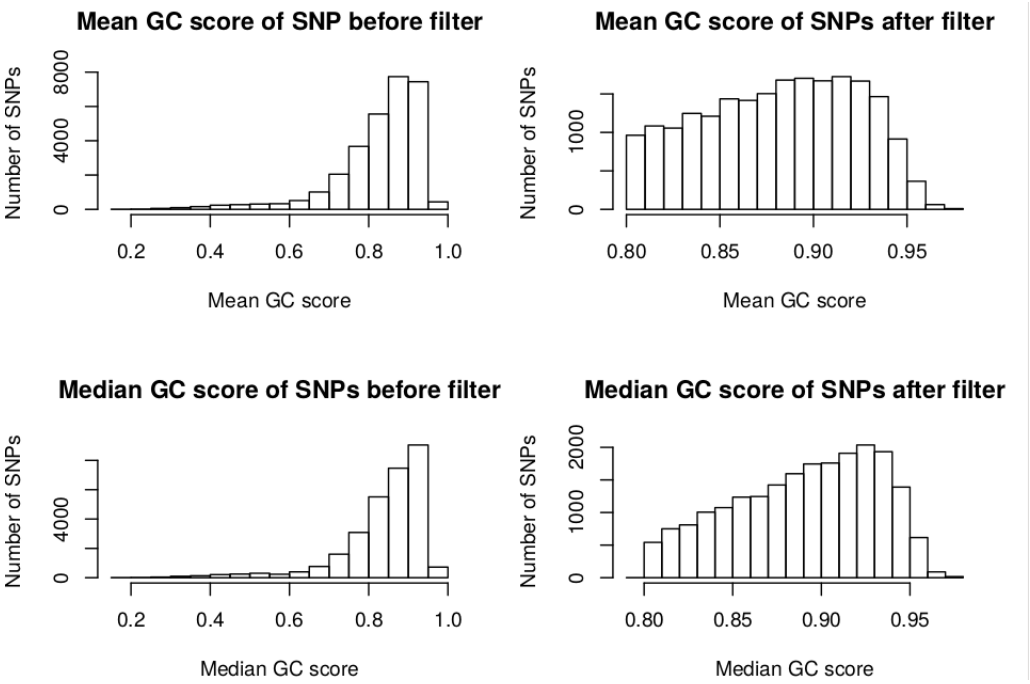
The GC score is a measurement of the genotype quality of Illumina technology. Samples with a mean and median GC score smaller than the cutoff are filtered out. A list of the SNPs with mean and median greater than the cutoff (Figure 3-3) are filtered out. A plot of the number of SNPs, before and after the filter, in function of the SNP's mean and median GC

score over all samples is shown (Figure 3-4). The function `qualityScoreBySnp` calculates the mean and median GC score. The workflow identifies the SNPs with mean and median bellow cutoff and filters them out. The results are in the folder *stratification_analysis*.

**Figure 3-3:** SNPs with mean and median greater than GC score cutoff

| snpID | mean.quality | median.quality |
|---|---|---|
| 8 | 0.83 | 0.90 |
| 10 | 0.87 | 0.88 |
| 12 | 0.83 | 0.83 |
| 15 | 0.81 | 0.81 |

**Figure 3-4:** Number of SNPs, before and after the filter, in function of the SNP's mean and median GC score over all samples.




- Population stratification analysis

The workflow recodes the genotypes with the 2, 1, or 0 B alleles structure. It allows performing a multiple correspondence analysis (MCA) between two the genetic origins of

the individuals and the SNPs with the function `dudi.acm` of the package ade4 (Dray & Dufour, 2007b). MCA is a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative (Abdi & Williams, 2010). With the function `fviz_mca_ind` of the package factoextra (Kassambara & Mundt, 2017), a scatter diagram with the coordinates of each sample generated in the MCA is plotted (Figure 3-5). Then, the function `assocRegression` of GWASTools carry out an association test between the SNPs and the genetic origin using logistic regression. A list with the SNPs associated with the origin is exported (Figure 3-6). After that, a Manhattan plot with the p-values of the association is shown (Figure 3-7). The tool removes the associated SNPs, and the results are stored again with the filtered data.

**Figure 3-5:** Scatter plot of multiple correspondence analysis of individuals from genetic origins.

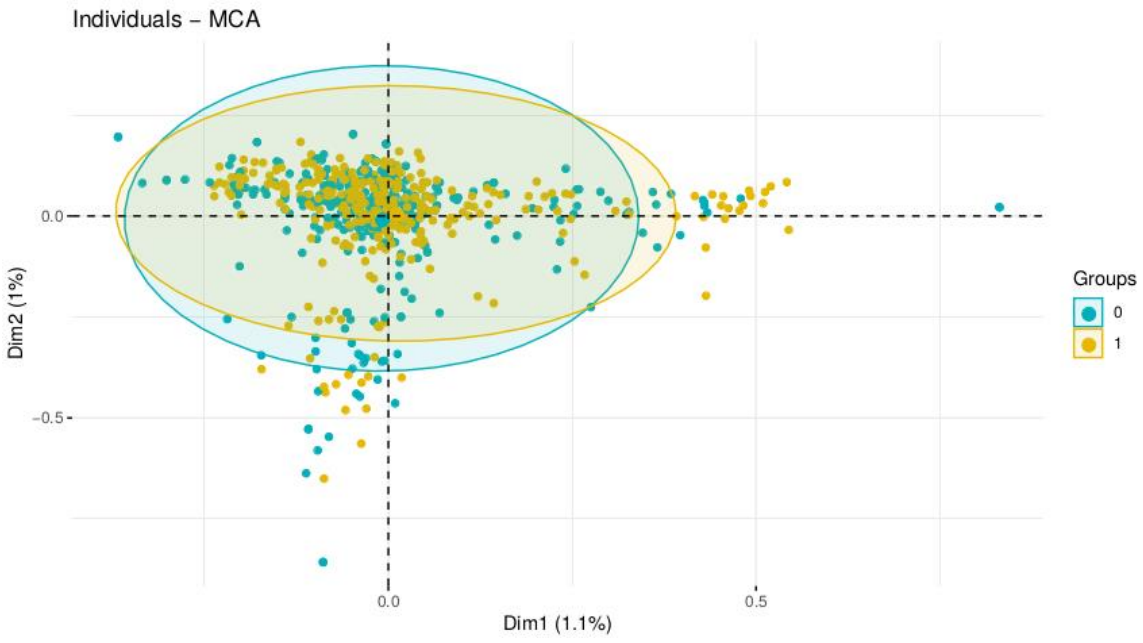
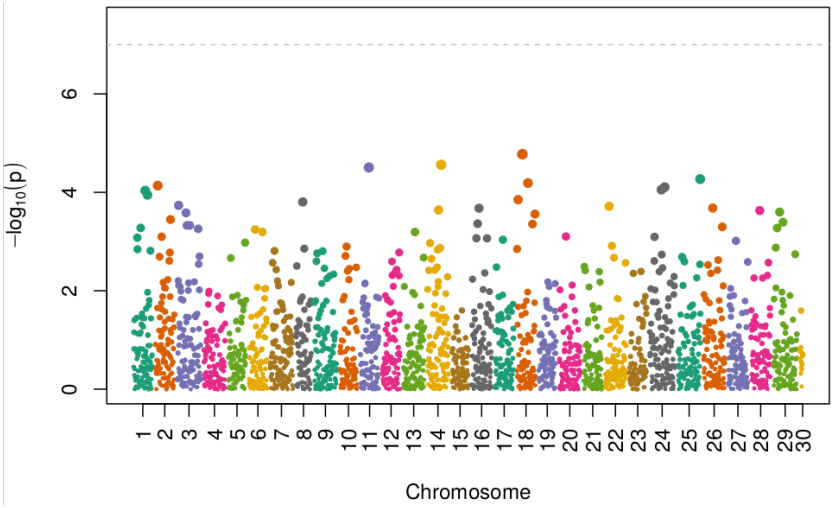


**Figure 3-6:** List of SNP associated with the origin.

```
snpID snpName
377 BTB-00003652
649 BovineHD0100010842
822 ARS-BFGL-NGS-38890
909 BTB-00026096
```

**Figure 3-7:** Manhattan plot of the association between SNPs and genetic origin.




- Hardy-Weinberg equilibrium

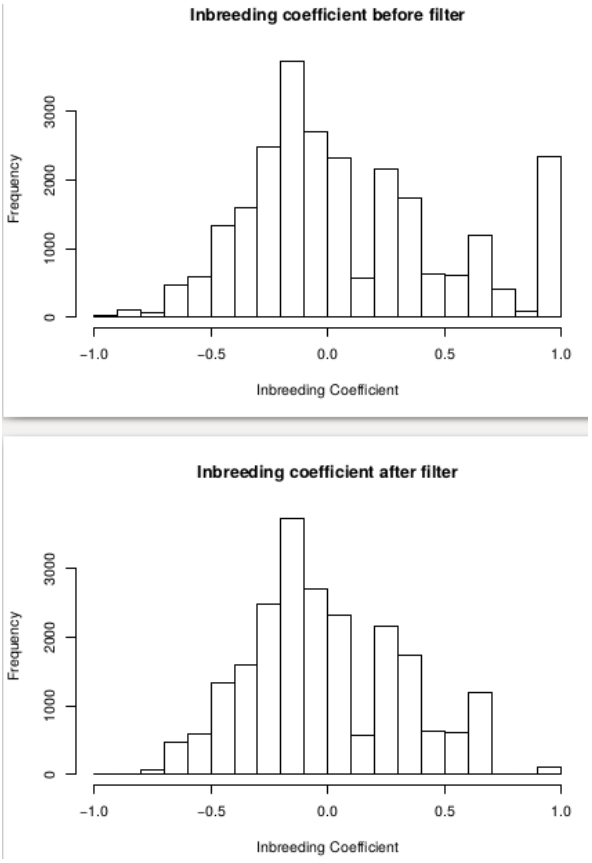
A population is in Hardy-Weinberg equilibrium (HWE) when the allele frequencies remain constant through the generations. In this section, the tool uses the Fisher's exact test to determine the departure of SNPs from HWE. The function used is `exactHWE` of the package GWASTools. Then, the code removes the SNPs with a score smaller than the cutoff. Files with HWE results before and after removing the SNPs are produced (Figure 3-8). With the HWE values per SNP, it can be calculated the inbreeding coefficient per SNP, which can show a possible population substructure. A distribution of inbreeding coefficients centered around 0 indicates there is most likely no significant population substructure (Gogarten et al., 2012). A plot of the number of SNPs per inbreeding coefficient value is used (Figure 3-9). Before performing the test, the workflow removes the founder individuals. HWE is done separately for autosomes and the X chromosome, and then, the results are merged. Below are examples of the results obtained.

**Figure 3-8:** Hardy-Weinberg equilibrium before (top) and after (bottom) the filter.

| snpID | chromosome | individuals_AA | individuals_AB | individuals_BB | Minor_allele_frequency | minor.allele | Inbreeding_coefficient | p-val | total_individuals |
|---|---|---|---|---|---|---|---|---|---|
| 287 | 1 | 3 | 1 | 7 | 0.318 | A | 0.790 | 0.01548 | 11 |
| 288 | 1 | 11 | 0 | 0 | 0.000 | B | NaN | NA | 11 |
| 289 | 1 | 3 | 3 | 5 | 0.409 | A | 0.436 | 0.21451 | 11 |
| 290 | 1 | 0 | 0 | 10 | 0.000 | A | NaN | NA | 10 |

| snpID | chromosome | individuals_AA | individuals_AB | individuals_BB | Minor_allele_frequency | minor.allele | Inbreeding_coefficient | p-val | total_individuals |
|---|---|---|---|---|---|---|---|---|---|
| 289 | 1 | 3 | 3 | 5 | 0.409 | A | 0.436 | 0.215 | 11 |
| 291 | 1 | 4 | 3 | 4 | 0.500 | A | 0.455 | 0.222 | 11 |
| 292 | 1 | 1 | 3 | 6 | 0.250 | A | 0.200 | 0.480 | 10 |
| 294 | 1 | 6 | 5 | 0 | 0.227 | B | -0.294 | 1.000 | 11 |

**Figure 3-9:** Inbreeding coefficient before (top) and after (bottom) the filter.




- Minor allele frequency analysis

The workflow calculates the minor allele frequency per each SNP with the function `exactHWE`. The tool removes the SNPs with a score smaller than the cutoff. The number of SNPs in function of their MAF before and after the filter are plotted (Figure 3-10). A text

file with MAF information of the SNPs with MAF greater than the cutoff is generated (Figure 3-11).

**Figure 3-10:** Number of SNPs in function of their MAF before and after the filter.

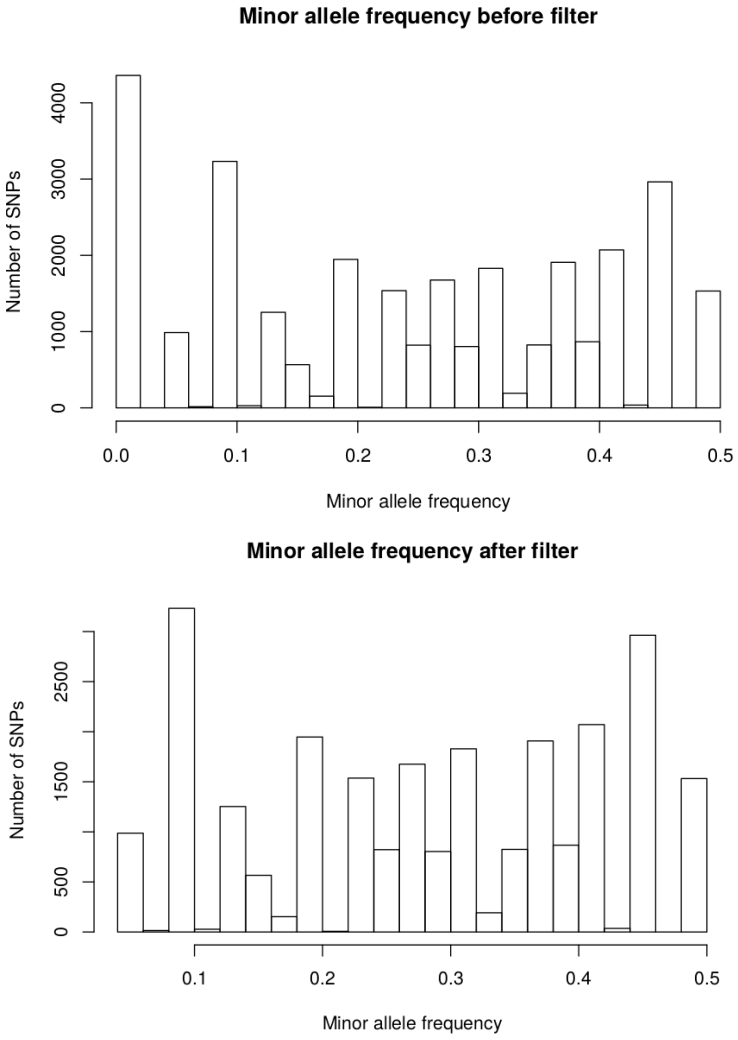



**Figure 3-11:** MAF information of the SNPs with MAF greater than the cutoff.

| snpID | chromosome | individuals_AA | individuals_AB | individuals_BB | Minor_allele_frequency | minor.allele | Inbreeding_coefficient | p-val | total_individuals |
|---|---|---|---|---|---|---|---|---|---|
| 287 | 1 | 3 | 1 | 7 | 0.318 | A | 0.790 | 0.01548 | 11 |
| 289 | 1 | 3 | 3 | 5 | 0.409 | A | 0.436 | 0.21451 | 11 |
| 291 | 1 | 4 | 3 | 4 | 0.500 | A | 0.455 | 0.22158 | 11 |
| 292 | 1 | 1 | 3 | 6 | 0.250 | A | 0.200 | 0.47988 | 10 |

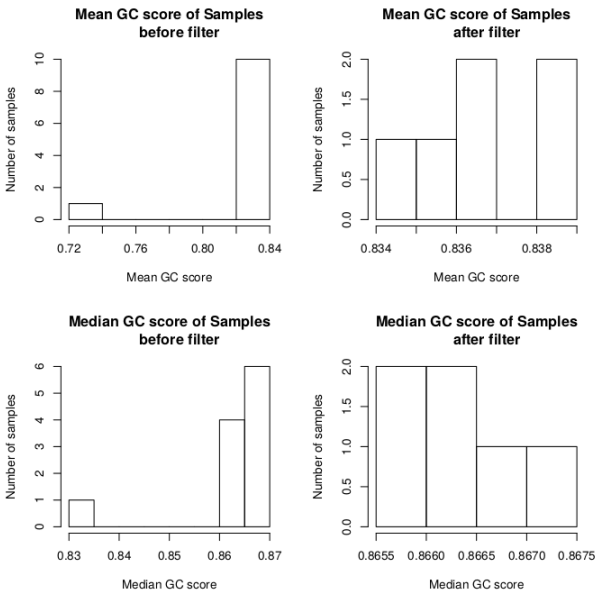
**Sample quality control**

The GC score per sample, the B Allele Frequency (BAF) and Log R Ratio (LRR), and the missingness and heterozygosity are analyzed. The tool uses the function `qualityScoreByScan` of GWASTools to calculate the GC score per sample. A list with the samples that have GC score mean and median greater than the cutoff (Figures 3-12 and 3-13) is generated, as well as plots of the number of samples before and after sample filtering

**Figure 3-12:** Samples with GC score mean and median greater than the cutoff.

| Sample | mean.quality | median.quality |
|---|---|---|
| 4 | 0.84 | 0.87 |
| 5 | 0.84 | 0.87 |
| 7 | 0.83 | 0.87 |
| 8 | 0.84 | 0.87 |
| 9 | 0.84 | 0.87 |
| 10 | 0.84 | 0.87 |

**Figure 3-13:** Number of samples before and after the filter out samples with GC score mean and median bellow cutoff.




At this point, the tool identifies the genomic region-samples pairs that have a BAF at least four standard deviations away from the mean of the BAF in the same region over all the

(Figure3-14 and Figure 3-15). Figure 3-15A shows an example of a genomic region – sample pair with less than four standard deviations of BAF from the mean of BAF of all samples in the same genomic region. Figure 3-15B shows an example of a genomic region – sample pair with more than four standard deviations of BAF from the mean of BAF of all samples in the same genomic region.

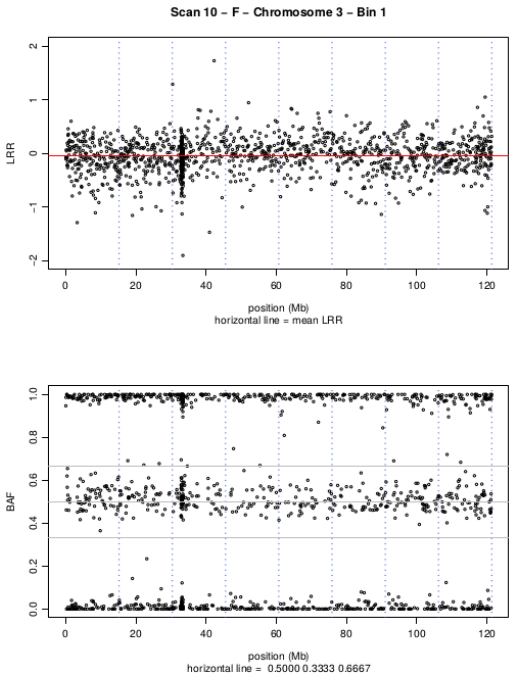
The function `sdByScanChromWindow` is applied to calculate the standard deviation of BAF at each window in each sample. The workflow calculates the mean of the BAF standard deviations in each region with the function `meanSdByChromWindow`. The function `findBAF` is used to identify samples with BAF standard deviation four times higher compared to other samples in a given region. These functions are from the package GWASTools.

**Figure 3-14:** List of the genomic region – sample pairs with more than four standard deviations of BAF from the mean of BAF of all samples in the same genomic region.

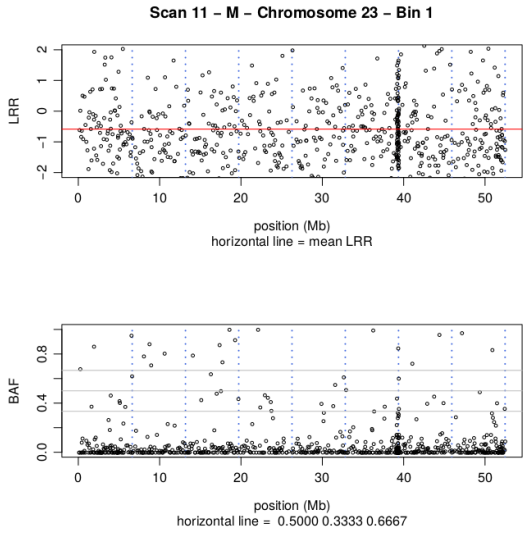
| scanID | chromosome | bin | sex |
|---|---|---|---|
| 11 | 23 | 1 | 2 |

**Figure 3-15:** BAF and LRR of the genomic region – sample pairs.

A




B



The code conducts tests for missingness and heterozygosity within samples. Samples with high heterozygosity may indicate a mixed sample. The tool identifies outliers regarding missingness and exports four plots: Missingness by chromosome (Figure 3-16), X chromosome missingness by sex (Figure 3-17), autosomal heterozygosity (Figure 3-18), and chromosome heterozygosity in females (Figure 3-19). Missingness is calculated with the function `missingGenotypeByScanChrom` of GWASTools. The workflow calculates the proportion of missingness with the `apply` function. The workflow calculates the heterozygosity by sample and chromosome with the function `hetByScanChrom` of GWASTools.

**Figure 3-16:** Missingness by chromosome.

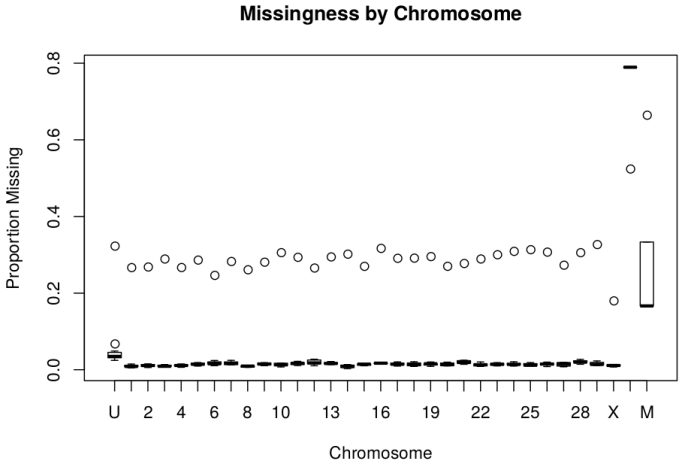



**Figure 3-17:** X Chromosome missingness by sex.

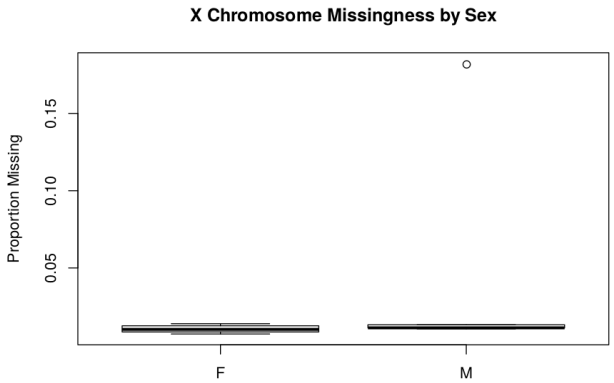


**Figure 3-18:** Autosomal heterozygosity.

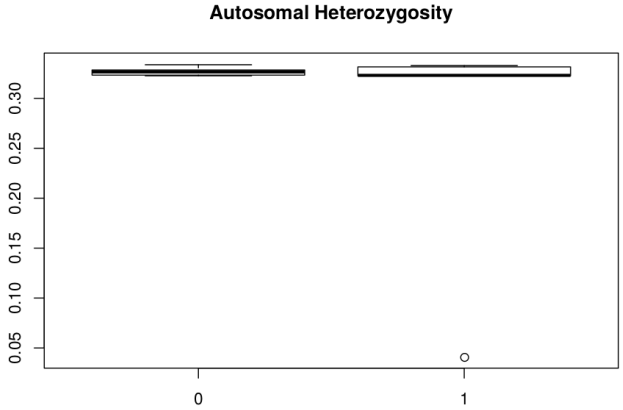
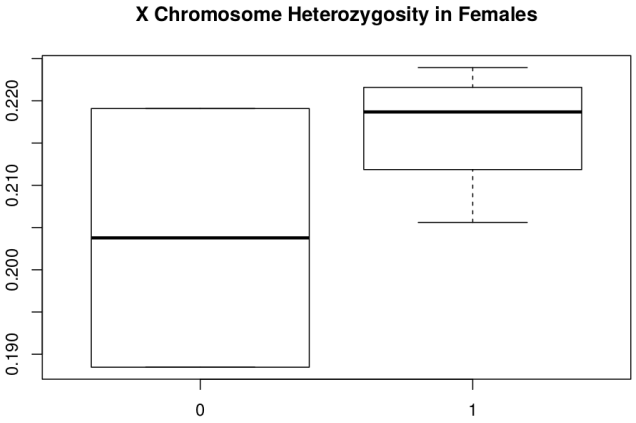



**Figure 3-19:** X Chromosome heterozygosity in females.




### Batch quality control

This section provides information about batch quality. A batch is a group of samples genotyped at the same time. In this section, the workflow performs the genotype quality per batch. The aim is to look for missing call rate (MCR) differences between batches. These analyses provide a view of the quality data to make decisions about batches. The tool creates three folders to store the results, one for the missing call before the filter, another for the missing call after the filter, and another that contains the last two. Additionally, the tool saves some results in the main folder of the batch quality control. These results are a plot with the number of samples per batch (Figure 3-20), a plot of the association between the mean MCR per batch and the number of samples per batch (Figure 3-21), a text with the mean MCR per batch (Figure 3-22), the genomic inflation factor per batch (Figure 3-23), and the association between batches and genetic origin (Figure 3-24).

The workflow shows results based on the analysis of the MCR per SNP and sample before and after filtering them out with a cutoff of 0.05. The results folders are *Missing_call_rate_before_filter* and *Missing_call_rate_after_filter*. Analyses of the MCR of the SNPs over all samples, MCR of the samples for all SNPs, MCR of the SNPs over samples whose MCR is greater than the cutoff, and MCR of the samples which MCR is greater than the cutoff are filtered out. After the filter, the results show analyses over all samples removing SNPs with high MCR, and for all SNPs removing samples with high MCR. In result number four (Figure 3-28), the proportion of SNPs per chromosome with MCR smaller than cutoff is not 1.0 because there are SNPs with an MCR greater than cutoff in non-remove samples. In the same sense, the results five (Figure 3-29) and nine (Figure 3-33) show result per SNPs removing high-MCR samples, and per sample removing high-MCR SNPs, respectively. The tool exports nine results before filter and the same nine after applying the filter. These results are listed below.

1. Number of samples, per sex, with missing calls in each SNP (Figure 3-25).
2. Number of samples by sex (Figure 3-26).
3. Fraction of missing calls per SNP (Figure 3-27).
4. Proportion of SNPs above MCR by chromosome (Figure 3-28).
5. Number of SNPs in function of MCR (Figure 3-29).
6. Missing counts per sample by chromosome (Figure 3-30).
7. Missing SNPs per chromosome (Figure 3-31).
8. Missing fraction per sample (Figure 3-32).
9. Number of samples in function of missing call rate (Figure 3-33).

Some examples of the results are shown below.

**Figure 3-20:** Number of samples per batch.

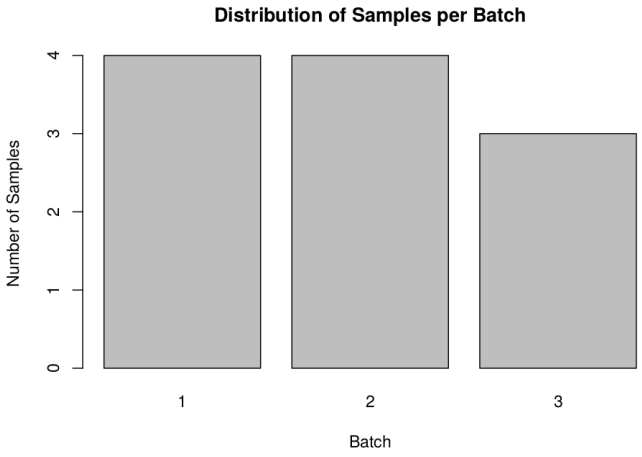



**Figure 3-21:** MCR per batch in function of the number of samples of the batch.

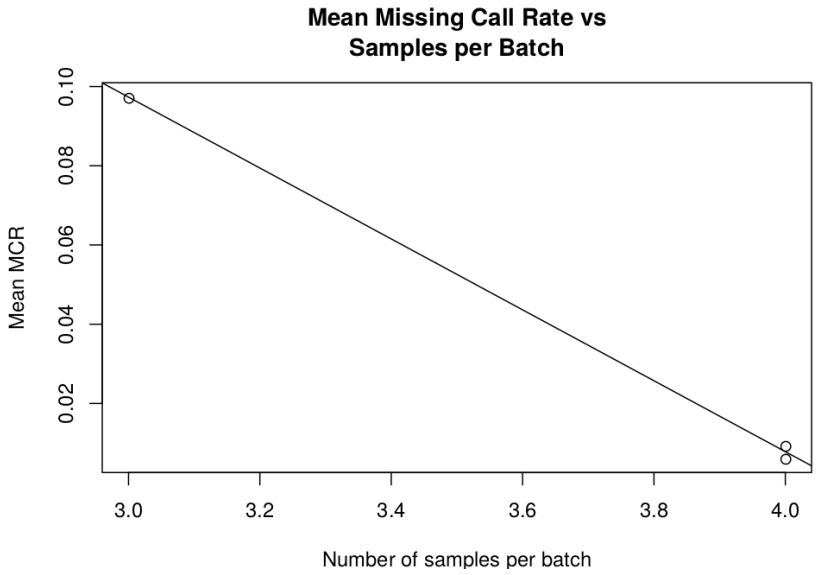



**Figure 3-22:** Mean missing call rate per batch.

| batch | mean_MCR |
|---|---|
| 1 | 0.0063 |
| 2 | 0.0081 |
| 3 | 0.2776 |

**Figure 3-23:** Genomic inflation factor per batch.

| Batch | Mean.chi.square | Genomic.inflation.factor |
|---|---|---|
| 1 | 0.43 | 0.22 |
| 2 | 0.48 | 0.25 |
| 3 | 1.11 | 0.57 |

**Figure 3-24:** Association between batches and genetic origin.

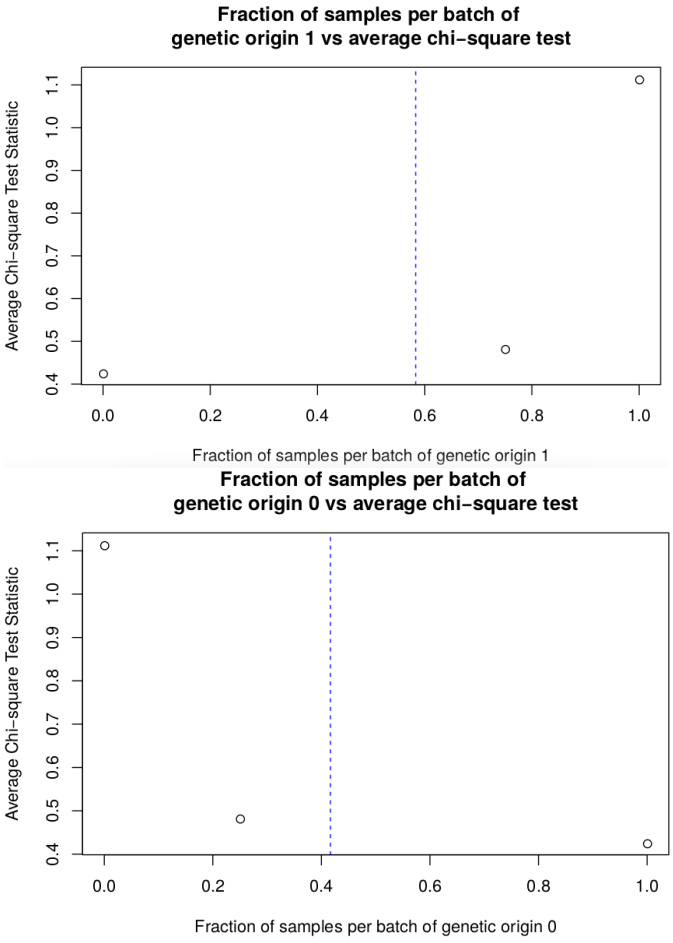



**Figure 3-25:** Missing counts by SNP and sex. M = male. F = female.

| snpID | snpName | M | F |
|---|---|---|---|
| 1 | Arachnomelia-BS | 0 | 0 |
| 2 | Arachnomelia-BS_2 | 0 | 0 |
| 3 | Arachnomelia-BS_3 | 0 | 1 |
| 4 | Arachnomelia-BS_4 | 0 | 0 |
| 5 | Arachnomelia-SM-F | 1 | 0 |

**Figure 3-26:** Samples by sex. M = Male. F = Female.

| | | | | | |
|---|---|---|---|---|---|
| M | 5 | Before filter | M | 4 | After filter |
| F | 6 | | F | 6 | |

**Figure 3-27:** The fraction of missing calls per SNP over all samples.

| snpID | snpName | missing.fraction |
|---|---|---|
| 1 | Arachnomelia-BS | 0.000 |
| 2 | Arachnomelia-BS_2 | 0.000 |
| 3 | Arachnomelia-BS_3 | 0.091 |
| 4 | Arachnomelia-BS_4 | 0.000 |
| 5 | Arachnomelia-SM-F | 0.091 |

Before filter

| snpID | snpName | missing.fraction |
|---|---|---|
| 1 | Arachnomelia-BS | 0.0 |
| 2 | Arachnomelia-BS_2 | 0.0 |
| 3 | Arachnomelia-BS_3 | 0.1 |
| 4 | Arachnomelia-BS_4 | 0.0 |
| 5 | Arachnomelia-SM-F | 0.0 |

After filter

**Figure 3-28:** The proportion of SNPs above MCR by chromosome.

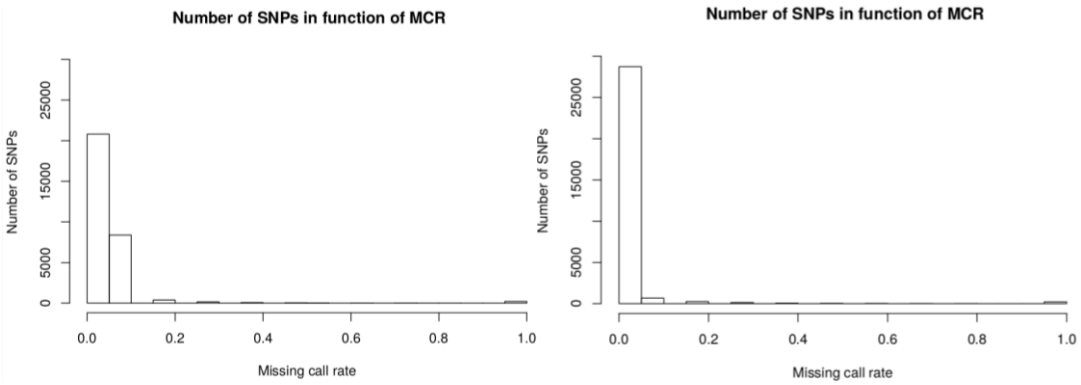
| Chromosome | Prop.of.SNPs.above.MCR |
|---|---|
| 1 | 0.704 |
| 2 | 0.704 |
| 3 | 0.687 |
| 4 | 0.715 |
| 5 | 0.685 |

Before filter

| Chromosome | Prop.of.SNPs.above.MCR |
|---|---|
| 1 | 0.96 |
| 2 | 0.96 |
| 3 | 0.96 |
| 4 | 0.97 |
| 5 | 0.95 |

After filter

**Figure 3-29:** Number of SNPs in function of MCR.




Before filter After filter

**Figure 3-30:** Missing counts per sample by chromosome.

| scanID | subjectID | Unknown | Chrom1 | Chrom2 | Chrom3 | Chrom4 | Chrom5 | Chrom6 | Chrom7 | Chrom8 | Chrom9 | Chrom10 | Chrom11 | Chrom12 | Chrom13 | Chrom14 | Chrom15 | Chrom16 | Chrom17 | Chrom18 | Chrom19 | Chrom20 | Chrom21 | Chrom22 | Chrom23 | Chrom24 | Chrom25 | Chrom26 | Chrom27 | Chrom28 | Chrom29 | XChrom | YChrom | Mitochondrial |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | TD01 | 5 | 11 | 7 | 14 | 11 | 16 | 12 | 20 | 8 | 9 | 6 | 11 | 20 | 7 | 10 | 7 | 6 | 6 | 10 | 5 | 11 | 14 | 6 | 8 | 8 | 6 | 7 | 4 | 4 | 9 | 8 | 1 | |
| 2 | TD02 | 3 | 13 | 16 | 8 | 6 | 19 | 7 | 15 | 8 | 11 | 13 | 5 | 17 | 8 | 6 | 5 | 7 | 7 | 7 | 8 | 11 | 13 | 2 | 3 | 5 | 8 | 0 | 6 | 6 | 6 | 8 | 8 | 0 |
| 3 | TD03 | 13 | 20 | 12 | 11 | 9 | 17 | 21 | 15 | 10 | 13 | 13 | 11 | 19 | 13 | 12 | 8 | 5 | 12 | 10 | 8 | 13 | 13 | 5 | 9 | 10 | 7 | 5 | 7 | 8 | 10 | 12 | 8 | 1 |
| 4 | TD04 | 2 | 10 | 11 | 12 | 5 | 12 | 16 | 9 | 10 | 13 | 6 | 8 | 6 | 7 | 4 | 7 | 6 | 11 | 3 | 7 | 4 | 8 | 2 | 7 | 5 | 4 | 4 | 5 | 3 | 4 | 11 | 8 | 0 |
| 5 | TD05 | 2 | 5 | 4 | 6 | 2 | 5 | 2 | 7 | 4 | 4 | 5 | 7 | 11 | 6 | 1 | 3 | 4 | 4 | 5 | 6 | 2 | 5 | 2 | 7 | 4 | 4 | 5 | 7 | 1 | 3 | 2 | 8 | 0 |
| 6 | TD06 | 7 | 9 | 13 | 11 | 5 | 14 | 9 | 11 | 8 | 9 | 12 | 15 | 15 | 10 | 11 | 7 | 7 | 8 | 10 | 8 | 8 | 9 | 9 | 4 | 5 | 5 | 7 | 5 | 3 | 9 | 8 | 8 | 1 |
| 7 | TD07 | 3 | 8 | 9 | 8 | 7 | 10 | 9 | 6 | 4 | 5 | 9 | 7 | 10 | 10 | 7 | 4 | 5 | 6 | 13 | 12 | 5 | 9 | 3 | 6 | 6 | 3 | 4 | 6 | 4 | 4 | 4 | 8 | 0 |
| 8 | TD08 | 0 | 4 | 5 | 6 | 0 | 6 | 7 | 6 | 5 | 5 | 2 | 4 | 4 | 5 | 1 | 4 | 6 | 3 | 2 | 5 | 5 | 5 | 1 | 2 | 4 | 3 | 7 | 8 | 0 | 9 | 0 | 3 | 0 |
| 9 | TD09 | 0 | 9 | 11 | 8 | 6 | 11 | 9 | 7 | 5 | 8 | 6 | 11 | 7 | 9 | 4 | 8 | 4 | 7 | 4 | 2 | 6 | 7 | 1 | 6 | 4 | 4 | 6 | 2 | 6 | 5 | 4 | 8 | 0 |
| 10 | TD10 | 5 | 17 | 17 | 7 | 4 | 10 | 16 | 13 | 8 | 11 | 8 | 17 | 10 | 8 | 7 | 5 | 6 | 5 | 3 | 5 | 7 | 12 | 4 | 6 | 6 | 4 | 4 | 1 | 5 | 2 | 10 | 8 | 0 |
| 11 | TD11 | 86 | 427 | 363 | 377 | 328 | 434 | 337 | 327 | 297 | 312 | 327 | 339 | 247 | 280 | 311 | 252 | 269 | 237 | 259 | 259 | 241 | 228 | 195 | 206 | 218 | 183 | 181 | 138 | 163 | 208 | 266 | 3 | 3 |

Before filter

| scanID | subjectID | Unknown | Chrom1 | Chrom2 | Chrom3 | Chrom4 | Chrom5 | Chrom6 | Chrom7 | Chrom8 | Chrom9 | Chrom10 | Chrom11 | Chrom12 | Chrom13 | Chrom14 | Chrom15 | Chrom16 | Chrom17 | Chrom18 | Chrom19 | Chrom20 | Chrom21 | Chrom22 | Chrom23 | Chrom24 | Chrom25 | Chrom26 | Chrom27 | Chrom28 | Chrom29 | XChrom | YChrom | Mitochondrial |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | TD01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | TD02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | TD03 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | TD04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | TD05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | TD06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | TD07 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | TD08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | TD09 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | TD10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | TD11 | 78 | 406 | 344 | 360 | 315 | 410 | 323 | 309 | 283 | 301 | 313 | 322 | 235 | 270 | 295 | 247 | 255 | 226 | 250 | 251 | 224 | 209 | 189 | 192 | 212 | 178 | 177 | 134 | 152 | 200 | 258 | 3 | 2 |

After filter

**Figure 3-31:** Missing SNPs per chromosome.

| chromosome | miss.snps.per.chr |
|---|---|
| U | 279 |
| 1 | 1598 |
| 2 | 1352 |
| 3 | 1303 |
| 4 | 1240 |
| 5 | 1528 |

Before filter

| chromosome | miss.snps.per.chr |
|---|---|
| U | 253 |
| 1 | 1534 |
| 2 | 1299 |
| 3 | 1258 |
| 4 | 1207 |
| 5 | 1464 |

After filter

**Figure 3-32:** Missing fraction per sample.

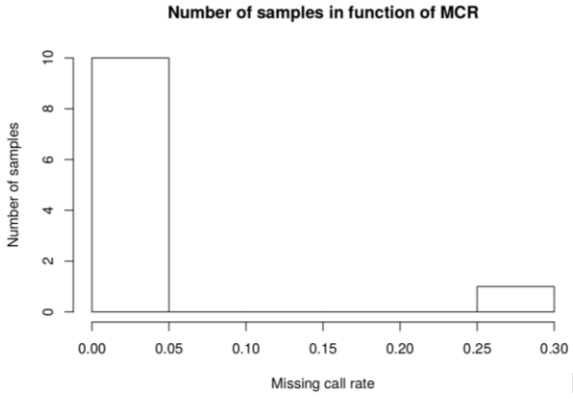
| scanID | subjectID | miss.missing.fraction |
|---|---|---|
| 1 | TD01 | 0.0095 |
| 2 | TD02 | 0.0089 |
| 3 | TD03 | 0.0118 |
| 4 | TD04 | 0.0077 |
| 5 | TD05 | 0.0044 |
| 6 | TD06 | 0.0094 |
| 7 | TD07 | 0.0069 |
| 8 | TD08 | 0.0042 |
| 9 | TD09 | 0.0063 |
| 10 | TD10 | 0.0081 |
| 11 | TD11 | 0.2776 |

Before filter

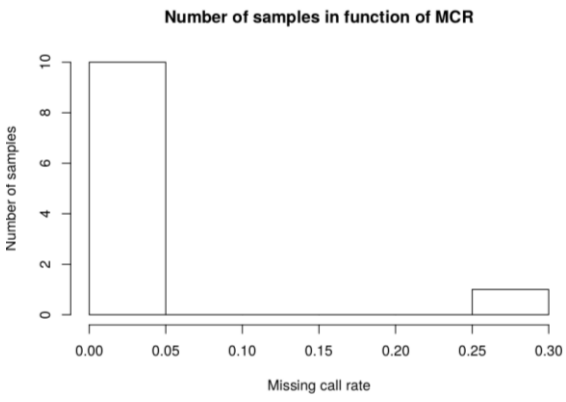
| scanID | subjectID | miss.missing.fraction |
|---|---|---|
| 1 | TD01 | 0.00 |
| 2 | TD02 | 0.00 |
| 3 | TD03 | 0.00 |
| 4 | TD04 | 0.00 |
| 5 | TD05 | 0.00 |
| 6 | TD06 | 0.00 |
| 7 | TD07 | 0.00 |
| 8 | TD08 | 0.00 |
| 9 | TD09 | 0.00 |
| 10 | TD10 | 0.00 |
| 11 | TD11 | 0.28 |

After filter

**Figure 3-33:** Number of samples in function of missing call rate.




Before filter




After filter

### 3.3.5 Association analysis

The association analysis is performed with the parameters provided by the user and using the function `assocRegression` of the package GWASTools (Gogarten et al., 2012). The analysis excludes the samples with low mean and median GC score. The tool adjusts the p-values with the multiple comparisons test defined by the user. The function is `p.adjust` of the package stats (R_Core_Team, 2018). A table with the association result (Figure 3-34), a Manhattan plot performed with the function `manhattanPlot` of GWASTools (Figure 3-35), and a list with the associated SNPs is generated (Figure 3-36).

**Figure 3-34.** Association result per SNP.

| snpID | chr | effect.allele | EAF | MAF | n | Est | SE | LL | UL | Wald.Stat | Wald.pval | P_adjust |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 289 | 1 | A | 0.350 | 0.350 | 10 | -3.3e-03 | 0.49 | -0.96067 | | 0.9541 | 4.5e-05 | 9.9e-01 | 1.0e+00 |
| 291 | 1 | A | 0.500 | 0.500 | 10 | -7.7e-01 | 0.33 | -1.41570 | | -0.1343 | 5.6e+00 | 1.8e-02 | 3.7e-01 |
| 292 | 1 | A | 0.278 | 0.278 | 9 | 2.6e-01 | 0.60 | -0.92552 | | 1.4360 | 1.8e-01 | 6.7e-01 | 1.0e+00 |
| 294 | 1 | B | 0.250 | 0.250 | 10 | 1.9e+00 | 0.37 | 1.14560 | | 2.6144 | 2.5e+01 | 5.2e-07 | 4.0e-04 |

**Figure 3-35:** Manhattan plot of the association analysis.

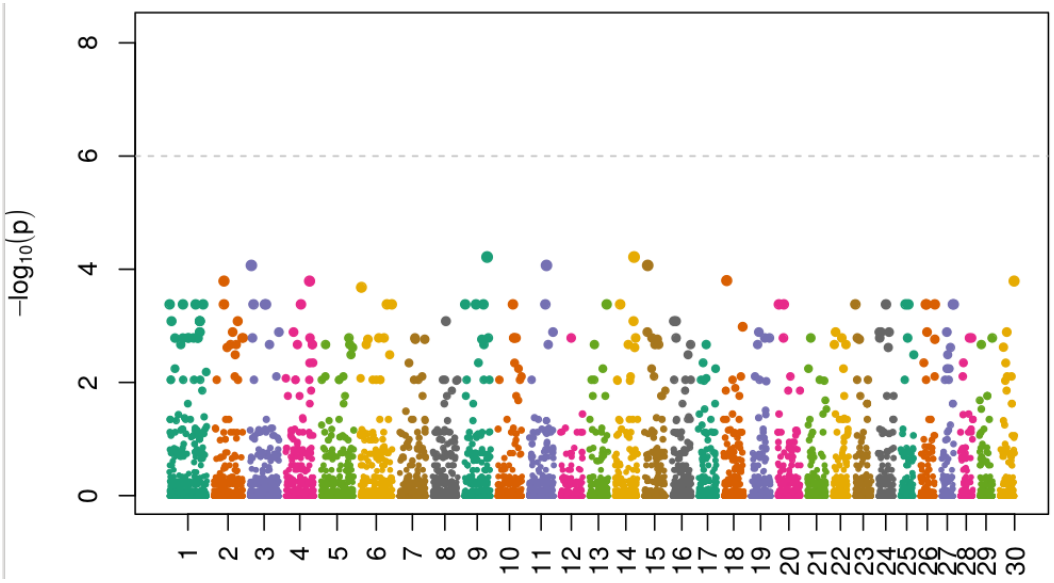



**Figure 3-36:** Information of the associated SNPs.

| snpID | chr | effect.allele | EAF | MAF | n | Est | SE | LL | UL | Wald.Stat | Wald.pval | P_adjust |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| snpName | chromosome | position | | alleleA | alleleB | BeadSetID | | tAA | tAB | tBB | rAA | rAB | rBB |

## 3.4 Description of module two – gene annotation

This module performs the annotation of one gene using the function `getGene` of the package mygene (Mark et al., 2018). The user creates the object gene with its Entrez or Ensembl gene identifier. With the function `print_annotation`, the tool formats the result, and exports it. Appendix A contains a result example working with the gene with Entrez id number 282659.

# 4. Genome-wide association study of meat tenderness in Colombian Creole Hair Sheep

**Authors:** B.J. Sepúlveda-Molina*, L. López-Klein**, M.F. Ariza***, S. Castro***, Y.T. Ortiz***, A.M. Amaya***, E.M. Rincón***

* Department of Systems and Industrial Engineering, National University of Colombia, Bogotá, 111321, Colombia. bjsepulvedam@unal.edu.co.
** Department of Statistics, National University of Colombia, Bogotá, 111321, Colombia.
*** Department of Animal Science, National University of Colombia, Bogotá, 111321, Colombia.

**Abstract.** Colombian Creole Hair Sheep (CCHS) has adapted to the low and middle tropical climate conditions of Colombia. This breed has the potential to be a genetic base of sheep meat production programs to compete in the international lamb market. In this market, the quality standards demanded a shire force of 5.0 Kg measured with the Warner-Bratzlet Shear Force (WBSF) test. The objective of the present work was to identify single nucleotide polymorphisms (SNPs) associated with the tenderness of roasted muscle *longissimus dorsi* of CCHS. 41 males and 62 females, from two geographical origins of Colombia, were genotyped with the OvineSNP50 BeadChip of Illumina. Mean, standard deviation, and ANCOVA analyses of WBSF was conducted in R. Data preprocessing, quality control, association analysis, and gene annotation were conducted with the tool Diploid-GWAS. Association analysis was performed with linear regression with sex and age as covariates, the statistical significance was evaluated with Wald test, and the p-values were adjusted with the False Discovery Rate test. Quality control included analyses of GenCall, minor allele frequency, Hardy-Weinberg equilibrium, and population stratification. 25,842 SNP were removed due to low quality. There was no population stratification. WBSF was 3.8 ± 0.98 kg, which is consistent with reported values in hair ovine breeds on similar

conditions. The interaction between sex and age affected the tenderness, and for that reason, they were used as covariates in the association analysis. A SNP was detected as associated with WBSF on chromosome 26 on the gene Teneurin transmembrane protein 3 (TENM3). TENM3 protein has two domains with functions associated with meat tenderness, the Epidermal growth factor (EGF) - like domain and the Carboxypeptidase-like regulatory domain.

**Keywords** — *genome-wide association study*, *single nucleotide polymorphism, sheep meat, mutton, meat tenderness, longissimus dorsi.*

## 4.1 Introduction

The Colombian sheep value chain is relatively young; it has been historically marginal, and its development depends on a higher participation on the lamb market (Buelvas & Pineda, 2008; Castellanos, Rodríguez, Toro, & Luengas, 2010). The lamb per capita consumption in Colombia is 310 g, representing the 0.4% of national meat intake, the lowest of all species and has had an annual decrease because the market cannot afford the price (Espinal, Martínez Covaleda, & Amézquita, 2006). Nevertheless, Colombia has a significant sheep number that could be used to export lamb. The international market pays around 4 US billion dollars each year for lamb. One of the principal importers is the United States (AgMRC, 2018). The leading exporters can reduce their offer by the effect of global warming in their production (Gowane et al., 2017). CCHS can supply part of this offer thanks to its adaptation to the tropical climate. The geographical proximity between Colombia and the USA represents an opportunity for Colombia to export mutton to the US. Additionally, other potential markets are the European Union, the United Kingdom, and México because they are importers and have commercial agreements with Colombia (ww.tlc.gov.co; G-3, 1994; Reina & Oviedo, 2011).

Competing in the world market requires to produce meat with high-quality standards. This quality depends on its development in meat by-products industry, as well as its nutritional level, and sensorial acceptability (Hamill, Marcos, Rai, & Mullen, 2012). Regarding sensorial acceptability, tenderness is generally considered the most influential variable and

day after day the customers pay higher prices for more tender meat (Aaslyng, Kerry, Ledward, & others, 2009; Goodson et al., 2002; Huffman et al., 1996; Lusk, Fox, Schroeder, Mintert, & Koohmaraie, 2001; Miller, Carr, Ramsey, Crockett, & Hoover, 2001; Rodas-González, Huerta-Leidenz, Jerez-Timaure, & Miller, 2009; Schroeder, Riley, & Frasier, 2008; Schroeder, Ward, Mintert, & Peel, 1998; Smith, Casas, Rexroad, Kappes, & Keele, 2000). There have been developed some mechanical methods to predict human perception of meat tenderness (O'Diam, 2009). The direct and most accepted instrumental method for tenderness measuring is Warner-Bratzler shear force (WBSF) (O'Diam, 2009). WBSF measures in kilograms (kg) the maximum shear force applied to a sample of meat by a blade of a texturometer. This test consists of a blade with a triangular hole, and blunt edge, which is used to cut a meat sample. Then, a spring dynamometer measures the force applied (AMSA, 2016). World market considers tough sheep meat with a WBSF value above 5 kg (Bianchi, Garibotto, Feed, Bentancur, & Franco, 2006; Safari, Channon, Hopkins, Hall, & Van De Ven, 2002).

CCHS is adapted to the low and middle Colombian tropic environment (Pastrana & Calderón, 1996). This natural adaptation increases rusticity, fertility, and disease resistance (Andersson & Georges, 2004; Egito, Mariante, & Albuquerque, 2002). CCHS was compared with commercial and adapted breeds from Brazil, Uruguay and Colombia and showed one of the lowest birth weights (2.5 kg) that could favor the lambing ease, and one of the highest weights at 365 days of age and adult weights, 37.52 kg and 80.12 kg, respectively (Carneiro et al., 2010).

Genome-wide association studies (GWAS) have contributed to improving the tenderness in sheep meat. These studies aimed to identify single nucleotide polymorphisms (SNP) associated with differences in tenderness between individuals of the same genetic composition and under similar environmental conditions. Ortiz *et al.* (2015) searched for SNPs associated with tenderness in cooked CCHS meat. They used the OvinSNP50 chip of Illumina and found three associated markers: OAR3_130491628.1, OAR4_118954127.1, and s43296.1. Illumina chip contains 6 SNPs close to genes associated with the cutting force in other species: CAPN1 (OAR21: 47225725), CAPN2 (OAR12: 28694194), CAPN3 (OAR7: 39000331) and CAST (OAR5: 101742566, OAR5: 101792466 and OAR5: 101853472). However, it has not been possible to predict the cutting

force when doing GWAS studies (Byun, Zhou, & Hickford, 2008; Ortiz et al., 2015; H. Zhou, Byun, Frampton, Bickerstaffe, & Hickford, 2008; H. Zhou, Hickford, & Fang, 2007). For this reason, an independent validation was carried out doing a GWAS with 1252 animals in Australia. These researchers genotyped the animals with 182 previously identified SNPs (M. I. Knight et al., 2012) and identified 3 SNPs in the CAST gene (CAST_101781475, CAST_101783060, and CAST_101829736) associated with the cutting force at the fifth-day post-mortem ($p < 0.05$). They did not find associations between SNPs of CAPN1, CAPN2 or CAPN3 (Matthew I. Knight et al., 2014). Knight *et al.* (2012) identified two SNPs in CAST (CAST_101781475 and CAST_101841509) and one in CAPN2 (CAPN2_28667683) associated with the shear force at fifth day post-mortem.

The objective of this work was to identify SNPs associated with meat tenderness in the CCHS using a GWAS with the chip OvinSNP50 of Illumina.

## 4.2 Materials and methods

**Animals and genotypes**

This work analyzed 103 CCHS individuals, 41 males, and 62 females, between 4 and 12 months of age, produced in extensive production based in a grazing system, organized into two sacrifice groups and from two geographical origins: *Valles Interandinos* (VI) and *Piedemonte* (PDM). The animals were genotyped using the chip OvinSNP50 Beadchip of Illumina. This chip is an array of 54,241 SNP probes. The chip was developed by a collaboration of Illumina, AgResearch, Baylor UCSC, CSIRO, and the USDA as part of the International Sheep Genomics Consortium (www.illumina.com).

**Facilities and Slaughtering**

The team received the samples at the Instituto de Ciencia y Tecnología de Alimentos of the Universidad Nacional de Colombia. The samples come from animals slaughtered after 24 hours of fasting. The *longissimus dorsi* muscle was removed, divided into two rashers of 160 gr approximated, and vacuum packaged.

The variable recorded was the Warner-Bratzler Shear Force (WBSF) following the protocol of AMSA (2016). The samples were roasted in a “George Foreman” griddle until they had an internal temperature of 72°C registered with a digital thermometer “Tylor 9842”. Later, the rashers were cold down to room temperature (17-21°C). From each rasher, three cores of 3 cm in length and 1.27 cm in diameter parallel to the longitudinal orientation of the muscle fibers were obtained with a hand-held coring device. The strength and hardness measurements were obtained using the software Exponent Soft and a texture analyzer “TA XT plus” with a pre-test speed of 2.0 mm/s, the test speed of 2.0 mm/s, the post-test speed of 10.0 mm/s, and a distance of 30mm.

**Data preprocessing and quality control**

This work used the R-workflow Diploid-GWAS available at https://github.com/bojusemo/Diploid-GWAS (Sepúlveda-Molina & López-Kleine, n.d.). The genotype file had the following data obtained with the software GenomeStudio of Illumina: Consecutive number of SNP, SNP name, chromosome, position, GenCall score, a fraction of nucleotide in the population, and theta and R values. The animal file contained animal id, sex, age, slaughter, demographic origin, strength, and hardness.

Three quality control (QC) steps were performed using the tool Diploid-GWAS. This work removed the SNPs with a GenCall Score smaller than 0.7, minor allele frequency analysis (MAF) smaller than 0.01, and Hardy-Weinberg equilibrium (HWE) p-value cutoff smaller than 0.05. With the HWE values per SNP, it can be calculated the inbreeding coefficient per SNP, which can show a possible population substructure. A distribution of inbreeding coefficients centered around 0 indicates there is most likely no significant population substructure (Gogarten et al., 2012). There was calculated the inbreeding coefficient per SNP. Additionally, a population stratification analysis was performed. multiple correspondence analysis (MCA) was carried out. Concerning the samples, This work made a test to identify individuals with a GenCall score smaller than 0.7. This score cutoff was select because 0.7 usually report well-behaving genotypes (Illumina, 2005)..

**WBSF analysis**

There was calculated the arithmetic mean and the standard deviation of WBSF using the functions `mean` of R base and `sd` of the package stats (R_Core_Team, 2018; R-Team,

2013). An analysis of covariance (ANCOVA) was conducted to evaluate the effect of sex, geographical origin, and age on WBSF. The effect of sex was included as a categorical variable with (male and female) with fixed effects. Geographical origin was included a categorical variable with the levels *pie de monte* (PDM) and *valles interandinos* (VI), with fixed effects. The age was included as a covariate. The test was performed with the function `aov` of the package stats (R_Core_Team, 2018).

**Association analysis and gene annotation**

After the QC, this work conducted an association test between SNPs and WBSF with linear regression using in the Diploid-GWAS tool (Sepúlveda-Molina & López-Kleine, n.d.). The association test used sex and age as covariates and computes p-values using the Wald tests. The sex and age were chosen as covariates because ANCOVA showed that their interaction affected WBSF. Additionally, the age has determined meat tenderness in other studies (Hopkins, Hegarty, Walker, & Pethick, 2006). The p-values were adjusted using the Benjamini-Hochberg method that seeks to control the "false discovery rate" (FDR) (Benjamini & Hochberg, 1995). A “conditional FDR Manhattan plot” was generated using Diploid-GWAS. In this plot, the SNP with conditional –log10 FDR>2 (that is FDR<0.01) is shown with the name of the gene where is located. The gene where the SNP is located was identified with the Genome Data Viewer browser of the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/genome/gdv). With Diploid-GWAS, an annotation of this gene was carried out looking for information in the biological databases.

## 4.3 Results and discussion

**Data preprocessing and quality control**

25,842 SNPs were removed due to low quality. 19,049 SNPs had a mean and median GenCall score under the cutoff of 0.7. The number of SNPs in function of the GenCall before and after remove SNPs by their GenCall are shown in Figure 4-1. 4,892 SNPs had a MAF smaller than 0.01, and the number of SNPs in function in MAF before and after applying the MAF filter are in Figure 4-2. 13,041 SNPs had an HWE p-value smaller than 0.05.

Figure 4-3 shows the distribution of inbreeding coefficients before and after filtering out SNPs with HWE criterium. The number of SNP removed by low GenCall score was higher than expected, corresponding to 35% of the markers. Population stratification analysis showed that were no markers associated with the genetic origin (Figure 4-4). There were no samples removed by low quality, bassing on the GenCall cutoff of 0.7.

**Figure 4-1:** Distribution of the number of SNPs in function of GenCall score before and after the filter.

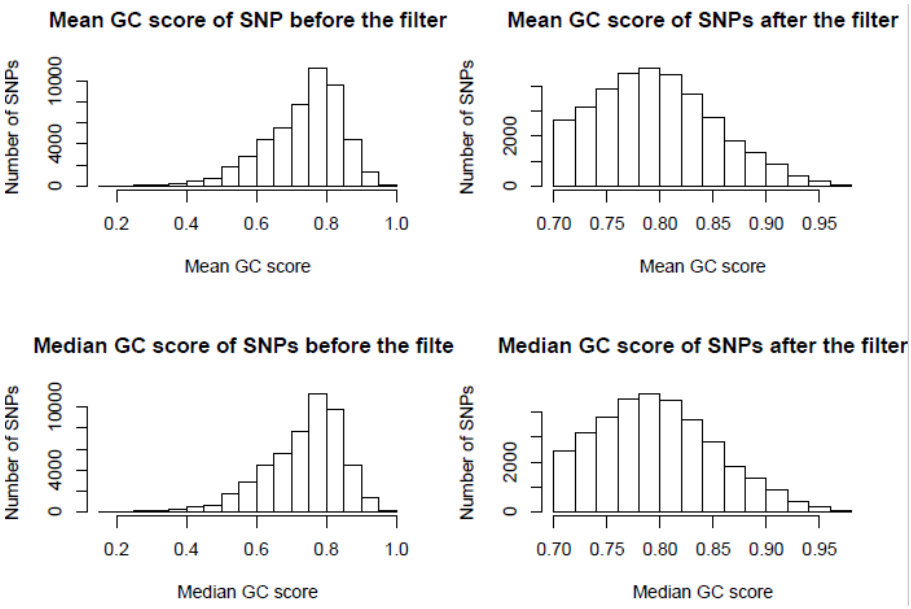



**Figure 4-2:** Minor allele frequency distribution before and after the filter.

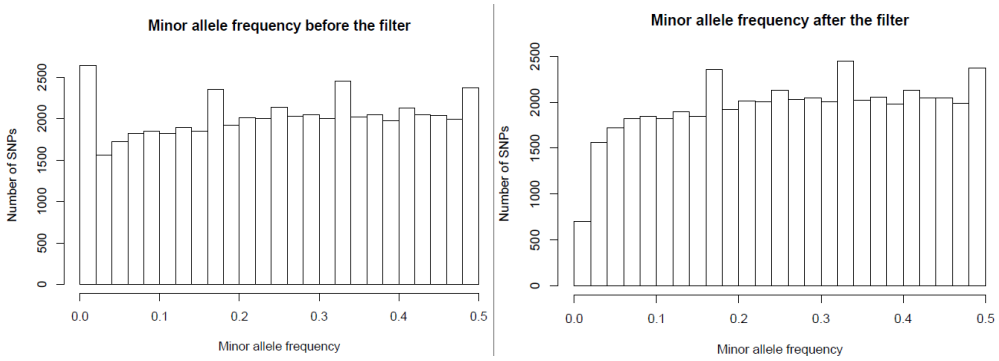


**Figure 4-3:** Inbreeding coefficient before and after the filter of HWE.

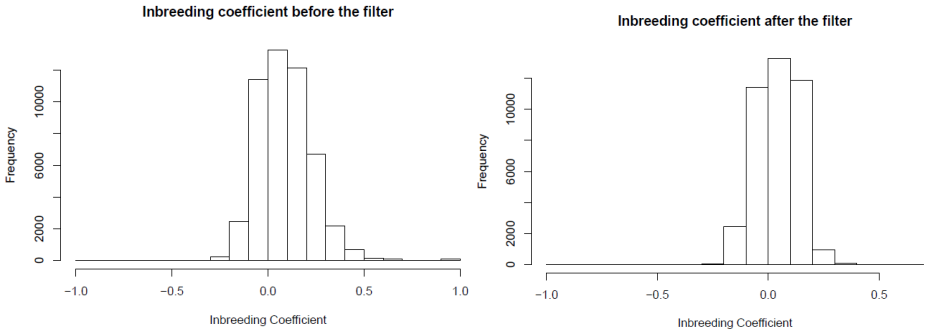
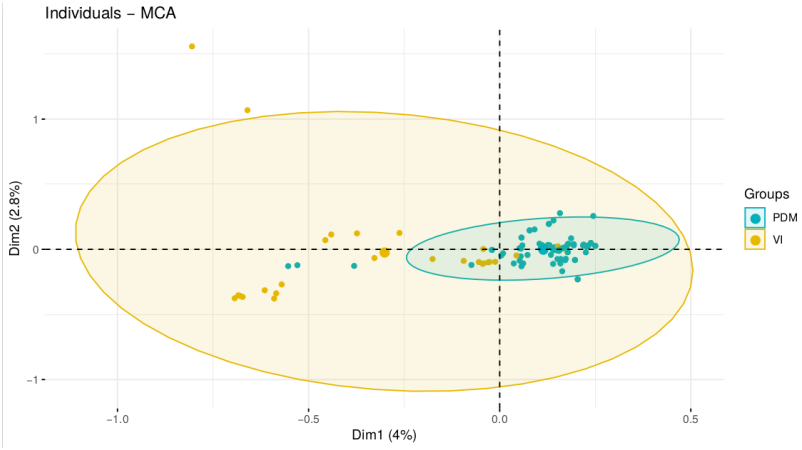



**Figure 4-4:** Scatter diagram of MCA coordinates.




**WBSF analysis**

The meat of CCHS showed a mean WBSF of 3.8 kg and a standard deviation of 0.98 kg. Burke and Apple (2007) measured the shear force of *longissimus* muscle of the hair sheep breeds Katahdin and St. Croix and the hair x wool breed Dorper in similar conditions to the present study and the results were similar. The shear force in this study was 3.8 kg, 3.8 kg, and 4.0 kg for Katahdin, St. Croix, and Dorper, respectively. Shackelford *et al.*(2012) reported the tenderness of the *longissimus* muscle for the hair breed Kathadin (4.7 kg), hair-wool breed Dorper (4.9 kg), and wool Finnsheep (4.4 kg), Romanov (4.7 kg), Rambouillet (5.1 kg), Suffolk (5.5 kg), Texel (4.9 kg), Dorset (5.2 kg), and Composite (5.7 kg) ($p<0.05$). These animals were fed with hey and concentrate, slaughtered before the

year of age, and their meat is less tender compared with CCHS. Another study reported that Dorper ram' descendant showed tender meat than Suffolk ram' descendant, being 2.8 kg and 3.98 kg, respectively ($p<0.05$) (Snowder & Duckett, 2003). The shear force of *longissimus dorsi* of lambs fed with commercial concentrate was smaller in the breeds St. Croix (2.21 kg) and crossbreeds 1/2 St. Croix + 1/2 Dorper (2.46 kg) and 1/2 Dorper + 1/4 de Suffolk + 1/4 Rambouillet (2.55 kg), than in the breed Suffolk (4.0 kg) and the crossbreeds 1/2 St. Croix + 1/4 Suffolk + 1/4 Rambouillet (3.96 kg), 1/2 St. Croix + 7/16 Suffolk + 1/16 Rambouillet-Dorset (4.97 kg), 11/16 Suffolk + 1/4 de Rambouillet + 1/16 Rambouillet-Dorset (4.92 kg) and 11/16 Suffolk + 5/16 Rambouillet (3.87 kg) ($p<0.05$) (Bunch et al., 2004).

Neither sex, geographical origin, nor age had a direct effect on the WBSF (Table 4-1). The same table shows a p-value for the interaction between sex and age of 0.026, which is significant at 0.05 level of significance. Then, the null hypothesis that 'there is no significant effect of interaction between sex and age on the WBSF' was rejected. Thus, it can be concluded that the WBSF depends on the interaction between sex and age.

**Table 4-1:** ANCOVA result. Effect of sex, geographical origin, and age on WBSF.

| | Df | Sum Sq | Mean Sq | F value | p |
|---|---|---|---|---|---|
| Sex | 1 | 2.297 | 2.297 | 2.425 | 0.123 |
| Geographical origin | 1 | 0.027 | 0.027 | 0.028 | 0.868 |
| Age | 1 | 2.065 | 2.066 | 2.18 | 0.143 |
| Sex x geographical origin | 1 | 1.023 | 1.023 | 1.08 | 0.302 |
| **Sex x age** | 1 | 4.85 | 4.85 | 5.12 | **0.026** |
| Geographical origin x age | 1 | 0.28 | 0.28 | 0.296 | 0.588 |
| Sex x geographical origin x age | 1 | 0.123 | 0.123 | 0.13 | 0.719 |
| Residuals | 91 | 86.219 | 0.948 | | |

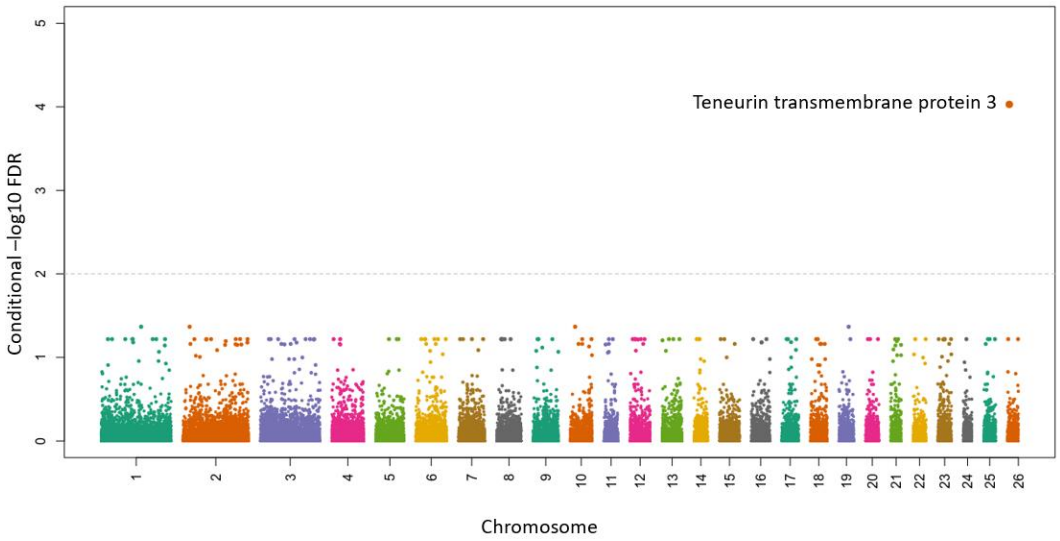
The age was included as a covariate.

**Association analysis and gene annotation**

The SNP OAR26_10469468.1 that is in the position 10469468 of chromosome 26 was detected as associated (Figure 4-5). The Wald p-value, p-value adjusted, and -log10 of the p-value were 3.49e-09, 9.34e-05, and 4.03, respectively. The SNP is in the gene Teneurin

transmembrane protein 3 (TENM3), a protein-coding gene. Appendix A stores the annotation result.

**Figure 4-5.** "Conditional FDR Manhattan plot". Conditional –log10 (FDR) for meat tenderness in Colombian Creole Hair Sheep. The SNP with conditional –log10 FDR>2 (that is FDR<0.01) is shown with name of the gene where is located.




TENM3 protein has two domains with functions associated with meat tenderness. These domains are Epidemial growth factor (EGF) - like domain (Interpro ID IPR000742), and Carboxypeptidase-like regulatory domain (Interpro ID IPR008969). The EGF links with its cell receptor on the cell surface and stimulates the activation of protein tyrosine kinase (Jura et al., 2009). Castro et al. (2017) reviewed and discussed how the tyrosine kinase might influence meat tenderness. A GWAS associated EGF-like domain with meat tenderness in Nellore cattle (Castro et al., 2017). Furthermore, the EGF affects the proportion of the type of muscle fibers and their areas, which determines the meat tenderness. The proportion of the types of fibers and their area affect the WBSF in bovines (Calkins, Dutson, Smith, Carpenter, & Davis, 1981). In muscle *longissimus dorsi* of pigs, the serum EGF was positively related to the percentage of type I muscle fiber ($r = 0.27$; $P < 0.015$), and negatively related to cross-sectional area percentage of type IIa fiber ($r = -0.21$; $P < 0.05$) (Ryu, Choi, Ko, & Kim, 2007). For its part, Carboxypeptidase B is a proteolytic enzyme presented in skeletal muscle cells that could facilitate postmortem changes associated with

the increase in meat tenderness (Koohmaraie, 1989). The present study is the first one that associate EGF-like and Carboxypeptidase-like regulatory domains with meat tenderness in sheep.

## 4.4 Conclusions

The meat tenderness of Colombian Creole Hair Sheep is competitive in the international market. These animals can contribute with its meat tenderness in breeding programs.

This work found an association between the SNP OAR26_10469468.1 and the meat shear force in the Colombian Creole Hair Sheep. This SNP is in the gene Teneurin transmembrane protein 3 (TENM3). TENM3 protein has two domains with functions associated with meat tenderness, the Epidemial growth factor (EGF) - like domain and Carboxypeptidase-like regulatory domain.

No SNP close to CAPN1, CAPN2, CAPN3 or CAST was associated with the meat shear force. This result coincides with previous studies. It is recommendable to work with customizing chips with more markers close to the mentioned genes.

## 4.5 Acknowledgment

This research was partially supported by the *Corporación Red* Especializada *de Centros de Investigación y Desarrollo Tecnológico del Sector Agropecuario de Colombia* (CENIRED) and the *Ministerio de Agricultura y Desarrollo Rural* of Colombia.

# 5. GWAS in Colombian population of Simmental cattle

**Authors**: B.J. Sepúlveda-Molina, L. López-Klein[1], L.M. Romero[2], Y.M. Gómez[2]

Department of Systems and Industrial Engineering, National University of Colombia, Bogotá, Colombia. bjsepulvedam@unal.edu.co.

[1] Department of Statistics, National University of Colombia, Bogotá, Colombia.

[2] Biotecnología y Genética S.A BIOTECGEN, Bogotá, Colombia.

**Abstract.** Cattle played a vital role in human evolution and currently is a critical portion of the Colombian rural economy. Genome-wide association studies (GWAS) have contributed to identifying single nucleotide polymorphisms (SNP) associated with economically important traits in the Simmental breed. This work aimed to identify SNPs associated with birth weight and 305-day milk yield in the Colombian Simmental population. The chip used was the GeneSeek GGP Bovine LD (30K). The quality control, association analysis, and gene annotation were performed with the open source tool Diplod-GWAS. The statistical model used was a linear regression per SNP. Quality control showed that genotypes had adequate quality in terms of GenCall, minor allele frequency, and Hardy-Weinberg equilibrium. Additionally, there was not identify population stratification. The SNP BovineHD4100012055 was associated with birth weight. The markers BovineHD1000024158, BovineHD0900002080, ARS-BFGL-NGS-7347, BovineHD2200017281, ARS-BFGL-NGS-66370, and BovineHD0500034987 were associated with 305-day milk yield.

**Keywords** — birth weight, genome-wide association study, milk production, pigmentation, single nucleotide polymorphism.

## 5.1 Introduction

The consumption of animal protein has played a fundamental role in the evolution of the human being and is an essential food for health and development of humanity (Bodwell & Anderson, 1986; Cottle & Kahn, 2014; Lombardi-Boccia, Lanzi, & Aguzzi, 2005). Cattle production contributes with 1.7% of the Colombian Gross Domestic Product (GDP), which represents 53% of the livestock GDP and 20% of the agricultural GDP, generating 7% of the national employment (FEDEGAN, 2011). Colombia produced, between 2002 and 2012, an average of 800 thousand tons/year of beef and 600 thousand tons/year of milk and its derivatives, the fourth largest producer in Latin America (FAO, 2013).

To compete in the international market, the Colombian Simmental producers' association, *Asociación Colombiana de Criadores de Ganado Simmental, Simbrah, Simmcebú y sus Cruces* (Asosimmental), has identified two traits and aims to improve them. Colombia signed trade agreements with exporting and importing countries such as the US, Canada, Mexico and Chile (www.tlc.gov.co). This situation obliges to Colombia to strengthen its presence in the domestic market and penetrate international markets (Beltrán et al., 2011). Optimizing the available resources is necessary. Therefore, it is essential to identify and improve the phenotypes of cattle breeds. Asosimmental has measured the birth weight and 305-day milk yield. Simmental cows tend to have calving difficulty and lower birth survival rates than other breeds because the calves are relatively heavy at birth (Comerford, Bertrand, Benyshek, & Johnson, 1987). Milk production is one of the most critical traits for Asosimmental because 60% of Colombian Simmental population is intended for milk production (Forero, 2014).

Genome-Wide Association Studies (GWAS) can be used to improve these traits. GWAS aim to identify associations between single nucleotide polymorphisms (SNPs) and traits and diseases in many species. GWAS in cattle, commonly use Illumina BeadChip technology given the low costs per genotype. GWAS have identified SNPs associated with multiple traits in Simmental (An et al., 2018; Song et al., 2016; Xia et al., 2017). Snelling et al. (2017) identified 293 SNPs associated with birth weight in 18 cattle breeds (Bonferroni-

corrected P < 0.05). The SNP rs29004488 in the leptin gene (LEP) and rs41974998 in milk fat globule-epidermal growth factor 8 protein gene (MFGE8) were associated with milk yield in the Italian Simmental population (P=0.043 and P=0.033, respectively). This study was conducted using a single marker regression model, which performs a regression for each marker.

Two processes are related to GWAS, a previous data quality control (QC) and the annotation of close genes to the SNPs associated with the phenotype. Before conducting GWAS, it is necessary to perform a data QC. QC includes genotype quality, as well as analyses of population stratification, Hardy-Weinberg equilibrium (HWE), and minor allele frequency (MAF). Genotype quality in Illumina BeadChip technology is reported as GenCall score (Oliphant et al., 2002). Population stratification can cause spurious associations (Somers et al., 2007). Between the techniques to identify population stratification is the multiple correspondence analysis between the SNPs and the genetic origin with the aim of found clusters of genetic origins (Cifuentes et al., n.d.). SNPs with Hardy-Weinberg imbalance may indicate genotyping errors, population stratification and even association with the phenotype (Turner et al., 2011). The exclusion of SNPs with low MAF avoids statistically unsupported associations (Gondro, Lee, et al., 2013).

This work aimed to identify SNPs associated with birth weight and 305-day milk yield in the Colombian Simmental population.

## 5.2 Materials and methods

**Data description**

This work is a pilot GWAS in the Colombian Simmental population. Data were obtained from a genomic evaluation of the breed performed by Asosimmental and the company Biotecnologia y Genetica S.A. The number of individuals analyzed, and the descriptive information of the traits analyzed in this study are shown in Table 1. In the genomic evaluation, the animals with phenotypic information and more descendants were selected to be genotyped. The DNA from the hair follicle was extracted using the phenol-chloroform standard method (Innis, Golfand, & White, 1990).

The samples were genotyped with the chip GeneSeek GGP Bovine LD (30K). The chip used contained 30,105 SNPs and was developed by GeneSeek, USDA-ARS and other collaborators using Illumina BeadChip technology. This chip includes evenly spaced and highly polymorphic SNPs with a mean gap of 89 kb, higher density marker placement at the telomeric region of the chromosomes for increased imputation accuracy. The imputation accuracy is higher than 99% in most well-characterized breeds included Simmental, the call rate success averages are above 99% and contains a large percentage of SNP overlap with other commercially available arrays including the original BovineSNP 50k of Illumina (www.neogen.com). It was assumed that the SNPs were in linkage equilibrium because the markers are expected to be linked to the same QTL alleles in distances less than 50 kb, as in the case of chips with more than 50,000 SNPs (Biegelmeyer, Gulias-Gomes, Caetano, Steibel, & Cardoso, 2016).

Birth weight is a continuous variable measured on the first day of life. 305-day milk is another continuous variable and measures the cow's milk yield from day 1 to day 305 of the lactation period (Kong et al., 2018). The number of lactations affects the 305-day milk yield (Ray, Halbach, & Armstrong, 1992). Based on it, 305-day milk yield for each lactation was treated as a new trait, creating three dependent variables. The phenotypic values are consistent with the reported in the breed (Comerford et al., 1987; Nistor et al., 2011).

**Table 5-1:** Traits included in the study.

| Trait | N | Mean | SD | CV% |
|---|---|---|---|---|
| Birth weight (kg) | 129 | 39.8 | 3.89 | 9.78 |
| First lactation 305-day milk yield | 64 | 3820 | 1414 | 37.02 |
| Second lactation 305-day milk yield | 80 | 5007 | 1667 | 33.29 |
| Third lactation 305-day milk yield | 62 | 4864 | 1585 | 32.59 |

**Quality control**

Simmental is used as a dual-purpose breed in Colombia. The sires of the animals registered in Asosimmental are mainly from Europe and the United States. Asosimmental recommends the use of sires from Europe when the goal is improving milk production traits, and bulls from the United States when the objective is to improve traits

associated with beef production (Forero, 2014). This situation could generate a population stratification.

For this reason, the population was divided according to the generic origin of the animals' sires. Then, to identify population stratification, a multiple correspondence analysis (MCA) was carried out between the SNPs and the genetic origin. The result was plotted in a scatter plot. These analyses were conducted utilizing Diploid-GWAS (Sepúlveda-Molina & López-Kleine, n.d.). With the same tool, there were removed SNPs with a GenCall Score smaller than 0.7, minor allele frequency analysis (MAF) lower than 0.01, and Hardy-Weinberg equilibrium (HWE) p-value cutoff smaller than 0.05. With the HWE values, it can be calculated the inbreeding coefficient per SNP, which can show a possible population substructure. A distribution of inbreeding coefficients centered around 0 indicates there is most likely no significant population substructure (Gogarten et al., 2012). QC also included removing samples with a GenCall score smaller than 0.7. The GenCall score cutoff followed the criterium of Illumina (Illumina, 2005).

**Association analysis and gene annotation**

There was performed a single marker regression GWAS. In this model, a randomly mating population without population stratification is assumed. According to Hayes (2013), the model has the following function:

$$y = Wb + Xg + e$$

Where $y$ is the phenotype vector, $W$ is a design matrix that assigns registers to fixed effects of the phenotypes, $b$ is a vector of fixed effects, $X$ is a design matrix that assigns registers to the effect of the marker, $g$ is the effect of the marker, and $e$ is a vector of random deviations $e_{ij} \sim N(0, \sigma_e^2)$, where $\sigma_e^2$ is the variance of the error.

Then, the Wald test was applied to evaluate the degree of contribution of each SNP to the phenotype. Wald test evaluates if $g$ is different of zero (null hypothesis) (Wasserman, 2004). The p-values were adjusted using the Benjamini-Hochberg method that seeks to control the "false discovery rate" (FDR) (Benjamini & Hochberg, 1995), and the result were called *conditional –log10 (FDR)*. A “conditional FDR Manhattan plot” per trait was performed. The manhattan plot cutoffs were conditional –log10 FDR>1.3 and 1.0 (that is, FDR<0.05 and 0.1, respectively).

The SNPs above the cutoffs were treated as associated with the traits. The closest gene to each of these markers was identified with the Genome Data Viewer browser of the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/genome/gdv). These genes were annotated with the tool Diploid-GWAS (Sepúlveda-Molina & López-Kleine, n.d.).

## 5.3 Results

**Quality control**

From the initial 31,105 SNPs, 27,144 SNPs had a mean and median GenCall score above the cutoff of 0.7. 25,235 SNPs had a MAF greater than 0.01. 27,113 SNPs had an HWE p-value greater than 0.05. Figure 1 shows the number of SNPs in function of inbreeding coefficients after filtering out SNPs that are not in HWE. This result indicates there is most likely no significant population substructure. Population stratification analysis showed that were no markers associated with the genetic origin (Figure 2). 21,968 SNPs had values greater than the cutoff of all parameters. The percentage of SNPs removed was similar in all chromosomes (25.5±5.8%). There was no a region with a high proportion of SNPs removed. There were no samples removed by low quality, based on the GenCall cutoff of 0.7. Supplementary material 1 includes the number of SNPs in function of the GenCall and MAF before and after CQ.

**Figure 5-1:** Number of SNPs in function of inbreeding coefficients after QC.

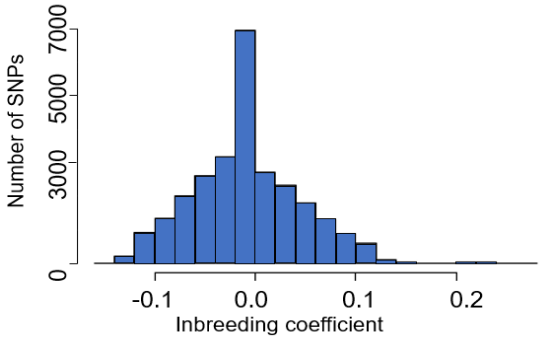
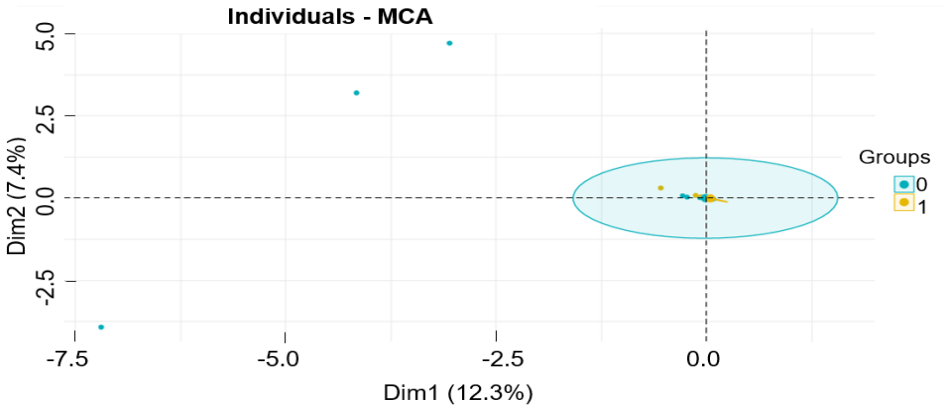


**Figure 5-2:** Scatter diagram of MCA coordinates. Different colors represent the genetic origin of the animal's sire, where 0 and 1 are Europe and United States origins, respectively.

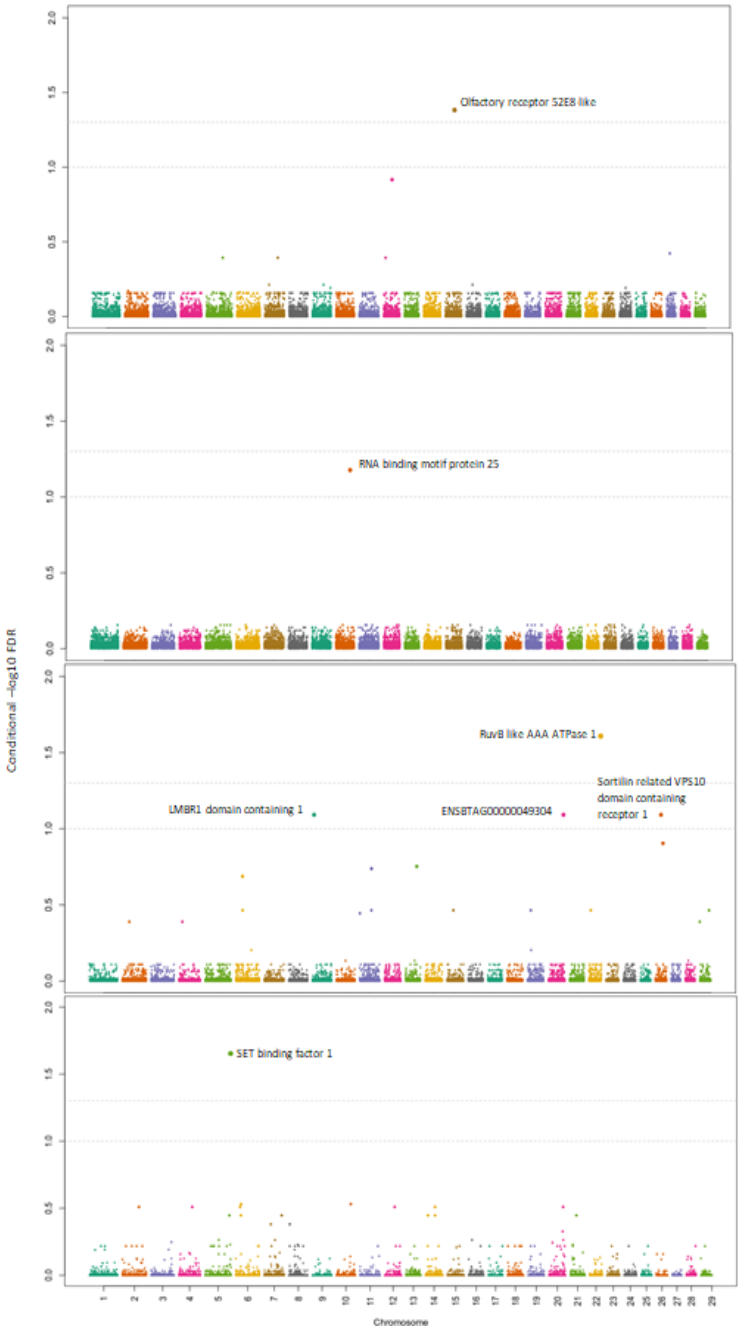



**Association analysis and annotation**

Seven SNPs were associated with the traits (Figure 3). Supplementary material 2 stores the annotation result. The SNP BovineHD4100012055 that is in the position 47195335 of chromosome 15 was associated with the birth weight (conditional –log10 (FDR) = 1.382). The closest gene to this marker is the olfactory receptor 52E8-like (Ensemble ID ENSBTAG00000000118). This gene coded for the protein with Ensembl ID ENSBTAP00000052030.

Six SNPs were associated with 305-day milk yield. The marker BovineHD1000024158 (position 84583993 and chromosome 10) was associated with the first lactation (conditional –log10 (FDR) = 1.178). The production in the second lactation was associated with the SNPs BovineHD0900002080, ARS-BFGL-NGS-7347, BovineHD2200017281, and ARS-BFGL-NGS-66370. The conditional –log10 (FDR) values were 1.608, 1.092, 1.092, and 1.092 respectively. These markers have the following position, respectively: 8709002 (chromosome 9), 67090770 (chromosome 20), 59530721 (chromosome 22), and 27942008 (chromosome 26). The SNP BovineHD0500034987 was associated with the trait in the third lactation, it is in chromosome 5 in position 119754719, and the conditional –log10 (FDR) was 1.653. The genes where are the SNPs or the closest genes to the markers are mentioned in the Manhattan plot. The protein domains are described in the supplementary material 2.

**Figure 5-3:** "Conditional FDR Manhattan plot". In descending order, conditional –log10 (FDR) values for birth weight, and first, second and third lactation 305-day milk yield. SNPs with conditional –log10 FDR>1.3 and 1.0 (that is, FDR<0.05 and 0.1, respectively) are shown with closets gene name or Ensembl gene ID.



## 5.4 Discussion and conclusions

Simmental cattle breed contributes with 150,000 liters of milk per day and with genetic material to produce meat in Colombia (ContextoGanadero, 2018). For this reason, the Simmental producer association, Asosimmental, have measured two phenotypes: birth weight and 305-day milk yield. We performed a GWAS to identify significant SNPs associated with these essential traits in Simmental cattle using a high-density SNP chip. Seven SNPs were associated with the two traits, and most SNPs were harbored on genes. Our results provide several novel markers associated with birth weight and 305-day milk yield in Simmental cattle.

This work is the first genome-wide association study on Colombian Simmental population, and some of the loci newly identified in this study may help to better DNA markers that determine increased beef and milk production in Simmental cattle. Further studies using a larger sample size will allow confirmation of the candidates identified in this study.

From the 30,105 SNPs analyzed, 21,968 SNPs had values greater than the cutoff of all QC parameters. There was no population stratification. The SNP BovineHD4100012055 was associated with birth weight. The closest gene to BovineHD4100012055 is the olfactory receptor 52E8-like (Ensemble ID ENSBTAG00000000118). This gene is a member of the three InterPro protein families: G protein-coupled receptor, rhodopsin-like; Olfactory receptor; and the G protein-coupled receptor (GPCR), rhodopsin-like, 7TM. None of these families have been reported as associated with birth weight in livestock species. However, GPCR has associated with birth weight in humans (Kovacs & Schöneberg, 2016). The markers BovineHD1000024158, BovineHD0900002080, ARS-BFGL-NGS-7347, BovineHD2200017281, ARS-BFGL-NGS-66370, and BovineHD0500034987 were associated with 305-day milk yield. Neither the closest genes nor the protein domains have been reported as associated with milk production.

## 5.5 Acknowledgment

Three institutions partially supported this research. The Colombian Simmental producers association, *Asociación Colombiana de Criadores de Ganado Simmental, Simbrah, Simmcebú y sus Cruces* (Assosimmental) provided the phenotypes. The

*Departamento Administrativo de Ciencia, Tecnología e Innovación* (Colciencias) financed the genotypes. The company *Biotecnologia y Genética S.A* executed this financing and provided these genotypes to the present work.

# 6. Conclusions and recommendations

## 6.1 Conclusions

A work methodology was standardized and organized in a workflow to perform the quality control, the association analysis between single nucleotide polymorphisms and phenotypes in ruminants and the subsequent biological contextualization of genes close to the SNPs associated with the phenotypes.

The work methodology was applied to the Colombian Creole Hair Sheep and Colombian Simmental cattle populations. The SNPs with low quality were removed. No samples were removed for low quality. There was no population stratification.

The SNP OAR26_10469468.1 was associated with the meat tenderness of Colombian Creole hair sheep. This SNP is in the gene Teneurin transmembrane protein 3 (TENM3). TENM3 protein has two domains with functions associated with meat tenderness, the Epidemial growth factor (EGF) - like domain and Carboxypeptidase-like regulatory domain.

The SNP BovineHD4100012055 was associated with birth weight. The closest gene to this SNP is the olfactory receptor 52E8-like, which is a member of the protein family G protein-coupled receptor (GPCR). GPCR has associated with birth weight in humans.

The markers BovineHD1000024158, BovineHD0900002080, ARS-BFGL-NGS-7347, BovineHD2200017281, ARS-BFGL-NGS-66370, and BovineHD0500034987, were associated with 305-day milk yield. Neither the closest genes nor the protein domains have been reported as associated with milk production.

## 6.2 Recommendations

- It is advisable to develop an R package with the workflow developed.
- It is recommended to use the tool developed with data of other species.
- It is advisable to improve the Hardy-Weinberg filter to avoid removing markers that are associated with the phenotype, as indicated by Turner *et al.* (2011).
- It is recommended to complement the tool developed with association analysis other than single marker regression.

# A. Appendix: compact disc

The disc contains the code and toy data described below.

- Module one QC and association - general structure
- Module one QC and association - GeneSeek structure
- Module two - annotation

Additionally, it also contains the following documents.

- Example of the annotation result
- Annotation of TENM3
- Supplementary materials 1 and 2 of chapter 5

# Bibliography


Aaslyng, M. D., Kerry, J. P., Ledward, D., & others. (2009). Trends in meat consumption and the need for fresh meat and meat products of improved quality. *Improving the Sensory and Nutritional Quality of Fresh Meat*, 3–18. Retrieved from http://www.cabdirect.org/abstracts/20093037369.html

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

AgMRC. (2018). International Lamb Profile. Retrieved from https://www.agmrc.org/commodities-products/livestock/lamb/international-lamb-profile

AMSA. (2016). *RESEARCH GUIDELINES FOR COOKERY, SENSORY EVALUATION, AND INSTRUMENTAL TENDERNESS MEASUREMENTS OF MEAT Second Edition*. Retrieved from http://www.meatscience.org/sensory

An, B., Xia, J., Chang, T., Wang, X., Miao, J., Xu, L., … Gao, H. (2018). Genome-wide association study identifies loci and candidate genes for internal organ weights in Simmental beef cattle. *Physiological Genomics*, *50*(7), 523–531. https://doi.org/10.1152/physiolgenomics.00022.2018

Andersson, L., & Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics*, *5*(3), 202–212.

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., … Dale, A. M. (2013). Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate. *PLoS Genetics*, *9*(4), e1003455. https://doi.org/10.1371/journal.pgen.1003455

Aulchenko, Y. S., de Koning, D.-J., & Haley, C. (2007). Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. *Genetics*, *177*(1), 577–585. https://doi.org/10.1534/genetics.107.075614

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: An R library for genome-wide association analysis. *Bioinformatics*, *23*(10), 1294–1296. https://doi.org/10.1093/bioinformatics/btm108

Ball, R. D. (2013). Designing a GWAS: Power, Sample Size, and Data Structure. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1019, pp. 37–98). https://doi.org/10.1007/978-1-62703-447-0_3

Beltrán, A., Ruiz, G., Santana, A., Estévez, L., Daza, B., Salamanca, O., … Arias, M. (2011). Seminario internacioanl del sector cárnico bovino, documento de memorias. Federación Colombiana de Ganaderos, FEDEGAN; Ministerio de Agricultura y Desarrollo Rural; Ministerio de Comercio, Industria y Turismo.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bianchi, G., Garibotto, G., Feed, O., Bentancur, O., & Franco, J. (2006). Efecto del peso al sacrificio sobre la calidad de la canal y de la carne de corderos Corriedale puros y cruza. *Archivos de Medicina Veterinaria*, *38*(2), 161–165. Retrieved from http://www.scielo.cl/scielo.php?pid=S0301-732X2006000200010&script=sci_arttext

Biegelmeyer, P., Gulias-Gomes, C. C., Caetano, A. R., Steibel, J. P., & Cardoso, F. F. (2016). Linkage disequilibrium, persistence of phase and effective population size estimates in Hereford and Braford cattle. *BMC Genetics*, *17*, 32. https://doi.org/10.1186/s12863-016-0339-8

Bodwell, C. E., & Anderson, B. A. (1986). Nutritional composition and value of meat and meat products. Academic Press. Retrieved from http://agris.fao.org/agris-search/search.do?recordID=US882887088

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. In *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* (pp. 3–62).

Bouwman, A. C., Visker, M. H., van Arendonk, J. A., Bovenhuis, H., Schennink, A., Stoop, W., … König, I. (2012). Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. *BMC Genetics*, *13*(1), 93. https://doi.org/10.1186/1471-2156-13-93

Buelvas, E., & Pineda, J. (2008). *Estudio de factibilidad para la creación de una empresa agropecuaria dedicada a la producción y comercialización de ovinos en pie en el municipio del El Roble, Sucre*. UNIVERSIDAD DE SUCRE .

Bunch, T. ., Evans, R. ., Wang, S., Brennand, C. ., Whittier, D. ., & Taylor, B. . (2004). Feed efficiency, growth rates, carcass evaluation, cholesterol level and sensory evaluation of lambs of various hair and wool sheep and their crosses. *Small Ruminant Research*, *52*(3), 239–245. https://doi.org/10.1016/j.smallrumres.2003.07.001

Burke, J. M., & Apple, J. K. (2007). Growth performance and carcass traits of forage-fed hair sheep wethers. *Small Ruminant Research*, *67*(2–3), 264–270. https://doi.org/10.1016/J.SMALLRUMRES.2005.10.014

Byun, S. O., Zhou, H., & Hickford, J. G. H. (2008). Haplotypic Diversity Within the Ovine Calpastatin (CAST) Gene. *Molecular Biotechnology*, *41*(2), 133–137. https://doi.org/10.1007/s12033-008-9103-2

Calkins, C. R., Dutson, T. R., Smith, G. C., Carpenter, Z. L., & Davis, G. W. (1981). Relationship of Fiber Type Composition to Marbling and Tenderness of Bovine Muscle. *Journal of Food Science*, *46*(3), 708–710. https://doi.org/10.1111/j.1365-2621.1981.tb15331.x

Carneiro, H., Louvandini, H., Paiva, S. R., Macedo, F., Mernies, B., McManus, C., … Lin, M. (2010). Morphological characterization of sheep breeds in Brazil, Uruguay and Colombia. *Small Ruminant Research*, *94*(1–3), 58–65. https://doi.org/10.1016/j.smallrumres.2010.07.001

Castellanos, J. G., Rodríguez, J. C., Toro, W. L., & Luengas, C. L. (2010). *Agenda prospectiva de investigación y desarrollo tecnológico para la cadena productiva cárnica ovino-caprina en Colombia*. (Giro Editores Ltda, Ed.). Bogotá, D.C., Colombia. Retrieved from http://docplayer.es/9854311-Agenda-prospectiva-de-investigacion-y-desarrollo-tecnologico-para-la-cadena-productiva-carnica-ovino-caprina-en-colombia.html

Castro, L. M., Rosa, G. J. M., Lopes, F. B., Regitano, L. C. A., Rosa, A. J. M., & Magnabosco, C. U. (2017). Genomewide association mapping and pathway analysis of meat tenderness in Polled Nellore cattle. *Journal of Animal Science*, *95*(5), 1945. https://doi.org/10.2527/jas2016.1348

Cifuentes, Y., Cortés, F., Franco, Á., & Niño, K. (n.d.). *Análisis de la ancestría como*

*herramienta para evaluar estratificación de la población en estudios de asociación del genoma completo (GWAS). !*

Comerford, J. W., Bertrand, J. K., Benyshek, L. L., & Johnson, M. H. (1987). Reproductive rates, birth weight, calving ease and 24-h calf survival in a four-breed diallel among Simmental, Limousin, Polled Hereford and Brahman beef cattle. *Journal of Animal Science*, *64*(1), 65–76. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3818492

ContextoGanadero. (2018). Raza Simmental aumentó 87,5 % producción de leche en 5 años. Contexto Ganadero.

Cottle, D., & Kahn, L. (2014). *Beef cattle production and trade*. CSRIO publishing.

Dominik, S. (2013). Descriptive statistics of data: understanding the data set and phenotypes of interest. *Methods in Molecular Biology (Clifton, N.J.)*, *1019*, 19–35. https://doi.org/10.1007/978-1-62703-447-0_2

Dray, S., & Dufour, A.-B. (2007a). The **ade4** Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, *22*(4), 1–20. https://doi.org/10.18637/jss.v022.i04

Dray, S., & Dufour, A.-B. (2007b). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, *22*(04), 1–20. Retrieved from https://econpapers.repec.org/article/jssjstsof/v_3a022_3ai04.htm

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*, 3439–3440.

Durinck, S., Spellman, P., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*, 1184–1191.

Egito, A. A., Mariante, A. D., & Albuquerque, M. S. M. (2002). Programa brasileiro de conservação de recursos genéticos animais. *Archivos de Zootecnia*, *51*(4), 39–52.

Espinal, C. F., Martínez Covaleda, H., & Amézquita, J. E. (2006). LA CADENA OVINOS Y CAPRINOS EN COLOMBIA. *DOCUMENTO DE TRABAJO No. 125* . Ministerio de Agricultura y Desarrollo Rural Observatorio Agrocadenas Colombia. Retrieved from http://agronet.gov.co/www/docs_agronet/20078611357_caracterizacion_ovinosycaprinos.pdf

Eusebi, P. G., González-Prendes, R., Quintanilla, R., Tibau, J., Cardoso, T. F., Clop, A., & Amills, M. (2017). A genome-wide association analysis for carcass traits in a commercial Duroc pig population. *Animal Genetics*, *48*(4), 466–469. https://doi.org/10.1111/age.12545

FAO. (2013). Estadísticas FAO.

FEDEGAN. (2011). Seminario internacional del sector cárnico bovino - documento de memorias.

Fernando, R. L., & Garrick, D. (2013). Bayesian Methods Applied to GWAS (pp. 237–274). https://doi.org/10.1007/978-1-62703-447-0_10

Forero, G. (2014). Simmental, una raza europea de 15 litros de leche diarios. *La Republica*. Retrieved from https://www.larepublica.co/archivo/simmental-una-raza-europea-de-15-litros-de-leche-diarios-2101719

Fortes, M. R. S., Reverter, A., Zhang, Y., Collis, E., Nagaraj, S. H., Jonsson, N. N., … Hawken, R. J. (2010). Association weight matrix for the genetic dissection of puberty in beef cattle. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(31), 13642–13647. https://doi.org/10.1073/pnas.1002044107

G-3. El tratado del grupo de los tres (G-3) entre los gobiernos de los Estados Unidos Mexicanos, la República de Colombia y la República de Venezuela (1994).

Gentleman, R. (2018a). annotate: Annotation for microarrays. R package version 1.60.0.

Gentleman, R. (2018b). Category: Category Analysis. R package version 2.48.0.

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., … Laurie, C. C. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics (Oxford, England)*, *28*(24), 3329–3331. https://doi.org/10.1093/bioinformatics/bts610

Gondro, C., Lee, S. H., Lee, H. K., & Porto-Neto, L. R. (2013). Quality Control for Genome-Wide Association Studies. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1019, pp. 129–147). https://doi.org/10.1007/978-1-62703-447-0_5

Gondro, C., Porto-Neto, L. R., & Lee, S. H. (2014). snpqc - an R pipeline for quality control of Illumina SNP genotyping array data. *Animal Genetics*, *45*(5), 758–761. https://doi.org/10.1111/age.12198

Gondro, C., van der Werf, J., & Hayes, B. (2013). *Genome-Wide Association Studies and Genomic Prediction*. (C. Gondro, J. van der Werf, & B. Hayes, Eds.) (Vol. 1019). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-447-0

Goodson, K. J., Morgan, W. W., Reagan, J. O., Gwartney, B. L., Courington, S. M., Wise, J. W., & Savell, J. W. (2002). Beef customer satisfaction: factors affecting consumer evaluations of clod steaks. *Journal of Animal Science*, *80*(2), 401–408. Retrieved from https://dl.sciencesocieties.org/publications/jas/abstracts/80/2/401

Gowane, G. R., Gadekar, Y. P., Prakash, V., Kadam, V., Chopra, A., & Prince, L. L. L. (2017). Climate Change Impact on Sheep Production: Growth, Milk, Wool, and Meat. In *Sheep Production Adapting to Climate Change* (pp. 31–69). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-4714-5_2

Grimm, D. G. (2015). *easyGWAS: An Integrated Computational Framework for Advanced Genome-Wide Association Studies*. https://doi.org/10.15496/PUBLIKATION-8322

Hamill, R., Marcos, B., Rai, D., & Mullen, A. (2012). Omics technologies for meat quality management. In *Omics Technologies: Tools for Food Science.* (pp. 249–282).

Hayes, B. (2013). Overview of statistical methods for genome-wide association studies (GWAS). *Genome-Wide Association Studies and Genomic*. Retrieved from http://link.springer.com/protocol/10.1007/978-1-62703-447-0_6

Hopkins, D. L., Hegarty, R. S., Walker, P. J., & Pethick, D. W. (2006). Relationship between animal age, intramuscular fat, cooking loss, pH, shear force and eating quality of aged meat from sheep. *Australian Journal of Experimental Agriculture*, *46*(7), 879. https://doi.org/10.1071/EA05311

Hudson, R. R. (2004). Linkage Disequilibrium and Recombination. In *Handbook of Statistical Genetics*. Chichester: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470022620.bbc23

Huffman, K. L., Miller, M. F., Hoover, L. C., Wu, C. K., Brittin, H. C., & Ramsey, C. B. (1996). Effect of beef tenderness on consumer satisfaction with steaks consumed in the home and restaurant. *Journal of Animal Science*, *74*(1), 91–97. Retrieved from https://dl.sciencesocieties.org/publications/jas/abstracts/74/1/91

Illumina. (2005). Illumina GenCall Data Analysis Software. Retrieved from https://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf

Innis, A., Golfand, D., & White, T. (1990). Sample preparation from blood, cells, and other fluids. In *HB Jovanovich, PCR Protocols: a Guide Methods and Applications* (pp.

356–367). Academic Press, San Diego.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Jura, N., Endres, N. F., Engel, K., Deindl, S., Das, R., Lamers, M. H., … Kuriyan, J. (2009). Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. *Cell*, *137*(7), 1293–1307. https://doi.org/10.1016/j.cell.2009.04.025

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., … Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. https://doi.org/10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Kassambara, A., & Mundt, F. (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5.

Knight, M. I., Daetwyler, H. D., Hayes, B. J., Hayden, M. J., Ball, A. J., Pethick, D. W., & McDonagh, M. B. (2012). Discovery and trait association of single nucleotide polymorphisms from gene regions of influence on meat tenderness and long-chain omega-3 fatty acid content in Australian lamb. *Animal Production Science*, *52*(7), 591–600. Retrieved from http://dx.doi.org/10.1071/AN11229

Knight, M. I., Daetwyler, H. D., Hayes, B. J., Hayden, M. J., Ball, A. J., Pethick, D. W., & McDonagh, M. B. (2014). An independent validation association study of carcass quality, shear force, intramuscular fat percentage and omega-3 polyunsaturated fatty acid content with gene markers in Australian lamb. *Meat Science*, *96*(2, Part B), 1025–1033. https://doi.org/10.1016/j.meatsci.2013.07.008

Kong, L., Li, J., Li, R., Zhao, X., Ma, Y., Sun, S., … Zhong, J. (2018). Estimation of 305-day milk yield from test-day records of Chinese Holstein cattle. *Journal of Applied Animal Research*, *46*(1), 791–797. https://doi.org/10.1080/09712119.2017.1403918

Koohmaraie, M. (1989). The role of endogenous proteases in meat tenderness. *Ars.Usda.Gov*. Retrieved from https://www.ars.usda.gov/ARSUserFiles/30400510/1988410089.pdf

Kovacs, P., & Schöneberg, T. (2016). The Relevance of Genomic Signatures at Adhesion

GPCR Loci in Humans. In *Handbook of experimental pharmacology* (Vol. 234, pp. 179–217). https://doi.org/10.1007/978-3-319-41523-9_9

Krzywinski, M., & Altman, N. (2014). Comparing samples—part II. *Nature Methods*, *11*(4), 355–356. https://doi.org/10.1038/nmeth.2900

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., … GENEVA Investigators. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591–602. https://doi.org/10.1002/gepi.20516

Laurie, C., Doheny, K., Mirel, D., & Pugh, E. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/gepi.20516/full

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835. https://doi.org/10.1038/nmeth.1681

Listgarten, J., Lippert, C., & Heckerman, D. (2013). FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics*, *45*(5), 470–471. https://doi.org/10.1038/ng.2620

Lombardi-Boccia, G., Lanzi, S., & Aguzzi, A. (2005). Aspects of meat quality: trace elements and B vitamins in raw and cooked meats. *Journal of Food Composition and Analysis*, *18*(1), 39–46. https://doi.org/10.1016/j.jfca.2003.10.007

Lusk, J. L., Fox, J. A., Schroeder, T. C., Mintert, J., & Koohmaraie, M. (2001). In-store valuation of steak tenderness. *American Journal of Agricultural Economics*, 539–550. Retrieved from http://www.jstor.org/stable/1245085

Mark, A., Thompson, R., Afrasiabi, C., & Wu, C. (2018). mygene: Access MyGene.Info_ services. R package version 1.16.2.

Marras, G., Rossoni, A., Schwarzenbacher, H., Biffani, S., Biscarini, F., & Nicolazzi, E. L. (2017). zanardi: an open-source pipeline for multiple-species genomic analysis of SNP array data. *Animal Genetics*, *48*(1), 121. https://doi.org/10.1111/age.12485

Miller, M. F., Carr, M. A., Ramsey, C. B., Crockett, K. L., & Hoover, L. C. (2001). Consumer thresholds for establishing the value of beef tenderness. *Journal of Animal Science*, *79*(12), 3062–3068. Retrieved from

https://dl.sciencesocieties.org/publications/jas/abstracts/79/12/3062

Nishimura, S., Watanabe, T., Mizoshita, K., Tatsuda, K., Fujita, T., Watanabe, N., … Takasuga, A. (2012). Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. *BMC Genetics*, *13*(1), 40. Retrieved from http://www.biomedcentral.com/1471-2156/13/40

Nistor, E., Bampidis, V. A., Pentea, M., Matiuti, M., Ciolac, V., & Adebambo, F. (2011). Genetic and phenotypic parameters for milk production traits in the first and second lactation in Romanian Simmental dairy cows. *Iranian Journal of Applied Animal Science*, *1*(4), 257–263. Retrieved from https://www.researchgate.net/publication/328739038_Genetic_and_phenotypic_parameters_for_milk_production_traits_in_the_first_and_second_lactation_in_Romanian_Simmental_dairy_cows

O'Diam, D. M. (2009). *Comparison of Slice Shear Force with Warner Bratzler Shear Force as Predictors of Consumer Panel Palatability Measures in Non-enhanced and Enhanced Pork Loin Chops*. Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=osu1236602515&disposition=attachment

Oliphant, A., Barker, D. L., Stuelpnagel, J. R., & Chee, M. S. (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *BioTechniques*, *Suppl*, 56–58, 60–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12083399

Ortiz, Y., Ariza, M., Castro, S., Ríos, M., & Sierra, L. (2015). Identificación genómica de snps asociados a terneza de la carne de ovino de pelo criollo colombiano. *Actas Iberoamericanas de Conservación Animal*, *6*, 388–397.

Pastrana, R., & Calderón, C. (1996). El ovino criollo colombiano. Los animales domésticos criollos y colombianos en la producción pecuaria nacional. *Instituto Colombiano Agropecuario – ICA.*

Pita, S., & Pértega, S. (n.d.). Asociación de variables cualitativas: test de Chi-cuadrado. Retrieved from https://www.fisterra.com/mbe/investiga/chi/chi.asp

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, *69*(1), 1–14. https://doi.org/10.1086/321275

Pritchard, J. K., Stephens, M., Rosenberg, N. A., Donnelly, P., Boehnke, M., Langefeld, C., … Ewens, W. (2000). Association mapping in structured populations. *American Journal of Human Genetics*, *67*(1), 170–181. https://doi.org/10.1086/302959

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007a). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007b). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

R_Core_Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

R-Team. (2013). R: A language and environment for statistical computing. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.5851&rep=rep1&type=pdf

Ray, D. E., Halbach, T. J., & Armstrong, D. V. (1992). Season and Lactation Number Effects on Milk Production and Reproduction of Dairy Cattle in Arizona. *Journal of Dairy Science*, *75*(11), 2976–2983. https://doi.org/10.3168/JDS.S0022-0302(92)78061-8

Reina, M., & Oviedo, S. (2011). Colombia y el TLC con la Unión Europea. *Friedrich Ebert Stiftung En Colombia, Fescol.*

Ren, X., Yang, G.-L., Peng, W.-F., Zhao, Y.-X., Zhang, M., Chen, Z.-H., … Li, M.-H. (2016). A genome-wide association study identifies a genomic region for the polycerate phenotype in sheep (Ovis aries). *Scientific Reports*, *6*(1), 21111. https://doi.org/10.1038/srep21111

Rentería, M. E., Cortes, A., & Medland, S. E. (2013). Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis (pp. 193–213). Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-447-0_8

Rodas-González, A., Huerta-Leidenz, N., Jerez-Timaure, N., & Miller, M. F. (2009). Establishing tenderness thresholds of Venezuelan beef steaks using consumer and

trained sensory panels. *Meat Science*, *83*(2), 218–223. Retrieved from http://www.sciencedirect.com/science/article/pii/S0309174009001284

Ryu, Y.-C., Choi, Y.-M., Ko, Y., & Kim, B.-C. (2007). Relationship between serum endocrine factors, histochemical characteristics of longissimus dorsi muscle and meat quality in pigs. *Journal of Muscle Foods*, *18*(1), 95–108. https://doi.org/10.1111/j.1745-4573.2007.00069.x

Safari, E., Channon, H. A., Hopkins, D. L., Hall, D. G., & Van De Ven, R. (2002). A national audit of retail lamb loin quality in Australia. *Meat Science*, *61*(3), 267–273. Retrieved from http://www.sciencedirect.com/science/article/pii/S0309174001001929

Schroeder, T. C., Riley, J. M., & Frasier, K. J. (2008). Economic Value of a Beef Tenderness-Based Fed Cattle Valuation System. *North American Institute for Beef Economic Research*. Retrieved from http://naiber.org/Publications/NAIBER/Fed.Cattle.Valuation.Sys.12pt.pdf

Schroeder, T. C., Ward, C. E., Mintert, J. R., & Peel, D. S. (1998). Value-based pricing of fed cattle: Challenges and research agenda. *Review of Agricultural Economics*, *20*(1), 125–134. Retrieved from http://aepp.oxfordjournals.org/content/20/1/125.short

Sepúlveda-Molina, B. J., & López-Kleine, L. (n.d.). Diploid-GWAS: An R workflow for quality control, GWAS, and annotation in diploid species.

Shackelford, S. D., Leymaster, K. A., Wheeler, T. L., & Koohmaraie, M. (2012). Effects of breed of sire on carcass composition and sensory traits of lamb1. *Journal of Animal Science*, *90*(11), 4131–4139. https://doi.org/10.2527/jas.2012-5219

Smith, T. P., Casas, E., Rexroad, 3rd CE, Kappes, S. M., & Keele, J. W. (2000). Bovine CAPN1 maps to a region of BTA29 containing a quantitative trait locus for meat tenderness. *Journal of Animal Science*, *78*(10), 2589–2594. Retrieved from https://dl.sciencesocieties.org/publications/jas/abstracts/78/10/2589

Snelling, W. M., Kachman, S. D., Bennett, G. L., Spangler, M. L., Kuehn, L. A., & Thallman, R. M. (2017). 197 Functional SNP associated with birth weight in independent populations identified with a permutation step added to GBLUP-GWAS. *Journal of Animal Science*, *95*(suppl_4), 97–98. https://doi.org/10.2527/asasann.2017.197

Snowder, G. D., & Duckett, S. K. (2003). Evaluation of the South African Dorper as a terminal sire breed for growth, carcass, and palatability characteristics. *Journal of Animal Science*, *81*(2), 368–375. https://doi.org/10.2527/2003.812368x

Somers, D. J., Banks, T., DePauw, R., Fox, S., Clarke, J., Pozniak, C., & McCartney, C. (2007). Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. *Genome*, *50*(6), 557–567. https://doi.org/10.1139/G07-031

Song, Y., Xu, L., Chen, Y., Zhang, L., Gao, H., Zhu, B., … Li, J. (2016). Genome-Wide Association Study Reveals the PLAG1 Gene for Knuckle, Biceps and Shank Weight in Simmental Beef Cattle. *PLOS ONE*, *11*(12), e0168316. https://doi.org/10.1371/journal.pone.0168316

Spielman, R. S., McGinnis, R. E., & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, *52*(3), 506–516. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8447318

Starmer, J. (2017). *StatQuest: FDR and the Benjamini-Hochberg Method clearly explained*. Retrieved from https://www.youtube.com/watch?v=K8LQSvtjcEo

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., & Aulchenko, Y. S. (2012). Rapid variance components–based method for whole-genome association analysis. *Nature Genetics*, *44*(10), 1166–1170. https://doi.org/10.1038/ng.2410

Tenenbaum, D. (2018). KEGGREST: Client-side REST access to KEGG. R package version 1.18.1.

Thomas, P. D. (2017). The Gene Ontology and the Meaning of Biological Function (pp. 15–24). Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-3743-1_2

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., … Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*, *Chapter 1*, Unit1.19. https://doi.org/10.1002/0471142905.hg0119s68

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Fernando, R. L., Vitezica, Z., … 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*, 134. https://doi.org/10.3389/fgene.2014.00134

Wasserman, L. (2004). *All of statistics : a concise course in statistical inference*. Springer.

Wiggans, G. R., Sonstegard, T. S., VanRaden, P. M., Matukumalli, L. K., Schnabel, R. D., Taylor, J. F., … Van Tassell, C. P. (2009). Selection of single-nucleotide

polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *Journal of Dairy Science*, *92*(7), 3431–3436. https://doi.org/10.3168/jds.2008-1758

Xia, J., Fan, H., Chang, T., Xu, L., Zhang, W., Song, Y., … Gao, H. (2017). Searching for new loci and candidate genes for economically important traits through gene-based association analysis of Simmental cattle. *Scientific Reports*, *7*(1), 42048. https://doi.org/10.1038/srep42048

Yu, Z., Demetriou, M., & Gillen, D. L. (2015). Genome-Wide Analysis of Gene-Gene and Gene-Environment Interactions Using Closed-Form Wald Tests. *Genetic Epidemiology*, *39*(6), 446–455. https://doi.org/10.1002/gepi.21907

Zaid, A., Hughes, H. G., Porceddu, E., & Nicholas, F. (2001). *Glossary of biotechnology for food and agriculture - A Revised and Augmented Edition of the Glossary of Biotechnology and Genetic Engineering*. Food and Agriculture Organization of the United Nations.

Zhang, C., Wang, Z., Bruce, H., Kemp, R. A., Charagu, P., Miar, Y., … Plastow, G. (2015). Genome-wide association studies (GWAS) identify a QTL close to PRKAG3 affecting meat pH and colour in crossbred commercial pigs. *BMC Genetics*, *16*(1), 33. https://doi.org/10.1186/s12863-015-0192-1

Zhang, H., Wang, Z., Wang, S., Li, H., Soller, M., Weigend, S., … Yang, N. (2012). Progress of genome wide association study in domestic animals. *Journal of Animal Science and Biotechnology*, *3*(1), 26. https://doi.org/10.1186/2049-1891-3-26

Zhang, Y., Guan, W., & Pan, W. (2013). Adjustment for Population Stratification via Principal Components in Association Analysis of Rare Variants. *Genetic Epidemiology*, *37*(1), 99. https://doi.org/10.1002/GEPI.21691

Zheng, X., Gogarten, S. M., Lawrence, M., Stilp, A., Conomos, M. P., Weir, B. S., … Levine, D. (2017). SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, *33*(15), 2251–2257. https://doi.org/10.1093/bioinformatics/btx145

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics (Oxford, England)*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Zhou, H., Byun, S. O., Frampton, C. M., Bickerstaffe, R., & Hickford, J. G. H. (2008). Lack

of association between CAST SNPs and meat tenderness in sheep. *Animal Genetics*, *39*(3), 331–332. https://doi.org/10.1111/j.1365-2052.2008.01720.x

Zhou, H., Hickford, J. G. H., & Fang, Q. (2007). Single nucleotide polymorphisms of the ovine calpain 3 (CAPN3) gene. *Molecular and Cellular Probes*, *21*(1), 78–79. https://doi.org/10.1016/j.mcp.2006.07.001

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310