Methodology for estimating association between categorical variables with application to Genome-wide association studies (GWAS)

Fabián Cortés Muñoz

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2019

Methodology for estimating association between categorical variables with application to Genome-wide association studies (GWAS)

Fabián Cortés Muñoz

Degree thesis presented as a partial requirement to qualify for the title of:
Doctor en Ciencias - Estadística

Doctoral Advisor:
Liliana López Kleine, M.Sc, Ph.D

Line of Research:
Bioestadística
Investigation Group:
Métodos en Bioestadística

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2019

A mi madre, por supuesto.

"En cierto modo, Aureliano José fue el hombre alto y moreno que durante medio siglo le anunció el rey de copas, y que como todos los enviados de las barajas llegó a su corazón cuando ya estaba marcado por el signo de la muerte. Ella lo vio en los naipes".

Taken from *Cien años de soledad*, Gabriel García Márquez (1967).

Acknowledgments

Special acknowledgement to the following persons:

Firstly, I would like to express my sincere gratitude to my thesis advisor Prof. Liliana López Klein, Ph.D. Thanks for believe in me since the beginning of this journey, for your continuous support as well as for your patience, motivation and immense knowledge.

A very special gratitude goes out to Miguel E. Otero Cadena, M.D., Vicerrector de Investigaciones at Universidad El Bosque. This personal accomplishment would not have been possible without your immense support, understanding and affection towards me. These few words are insufficient to thank the immense gratitude that I have for you.

I would also like to acknowledge to Andrés Felipe Hernández Bustos, S.B., for your contribution to the programming in `R`. Without his passionate participation and input, this proposal would never have given birth.

Luis Alejando Másmelas, thanks for being there when the doubts came on my mind over and over. Thanks for teaching me, support me and help me.

Finally, last but by no means least, I must express my very profound gratitude to my family: mom and Iván, thanks for your infinite love, understanding and support. If it were not for you both, I would have given up on my dreams. If it were not for you both, I would walk around like a dog without an owner. Merlina, Miranda and Patricio, thank you for always being there, no matter what happens. Thanks to Ronald, Francy, María Valeria, Luz Omaira and Luis Felipe.

Abstract

Several genomic data analysis contexts have a large number of statistical hypotheses, which are tested simultaneously. When the association between categorical phenotypes (i.e. healthy and not healthy) and Single Nucleotide Polymorphisms (SNPs) are assessed by applying statistical tests, the two key challenges to address are the following: which method is the best for using multiple testing and how to increase the statistical power after adjustment for multiple testing. In this association studies, a solid criterion obtained to consider its significant with high statistical power and without necessarily increasing the sample size is crucial. Numerous methods have been developed for addressing these limitations; they have improved type I and type II errors rates. The proposed methods are mainly based on changing the type for establishing the association and extending it to continuous traits.

Some of these statistical methods are very complex, which are difficult to use, specially for non-statisticians who usually obtain such data. Moreover, very few methods focused on developing a new statistical test for categorical data, which is the most common form of measuring phenotypical traits in humans and other organisms. By applying the maximum values of chi-square distribution as the test statistic, this study propose a new statistical test called Quotient C that allows testing associations between thousands of SNPs and a categorical trait.

In real datasets, Quotient C is observed to be less stringent criterion that allows the declaration of a large number of associations between SNPs and dichotomous outcomes in comparison with the classical methods used for correcting multiple testing, thus keeping the probability of incorrectly rejecting a true null hypothesis (type I error) equal or less than α. The proposed method has a lower type II error rate and a better statistical power than the following methods: Bonferroni, Holm, Hochberg and Benjamini and Hochberg.

Key words: Genome-wide association studies, multiple testing problem, categorical data, type I error, statistical power.

Resumen

En varios contextos de análisis de datos genómicos, una gran cantidad de hipótesis estadísticas se ponen a prueba al mismo tiempo sobre un mismo conjunto de datos. En Estudios de asociación del genoma completo (GWAS, por sus siglas en inglés), se evalúa la asociación entre fenotipos categóricos (como por ejemplo, sanos versus enfermos) y la presencia o ausencia de polimorfismos de nucleótido simples (SNPs, por sus siglas en inglés) mediante pruebas estadísticas. En éstos tipos de estudios, dos problemas principales se deben abor-

dar: qué método usar para la corrección por pruebas múltiples y cómo aumentar la potencia estadística después de realizar ajustes al error tipo I. La obtención de un criterio sólido para declarar asociaciones significativas con alto poder estadístico sin aumentar el tamaño de la muestra es crucial para estos estudios de asociación. Aademás los métodos propuestos deben ser fáciles de usar y comprender para los científicos que producen los datos.

Diversos métodos han sido desarrollados a fin de abordar tales limitaciones con importantes resultados en la disminución de los errores de tipo I y tipo II. Los métodos propuestos se basan principalmente en cambiar el tipo de prueba para establecer la asociación y extenderla a fenotipos continuos. Algunos de ellos son métodos muy complejos y son difíciles de usar para los no estadísticos, quienes son generalmente los que producen los datos. Además, muy pocos métodos se han centrado en desarrollar una nueva prueba estadística para datos categóricos, que es la forma más común de medir rasgos fenotípicos en humanos y otros organismos.

Una nueva prueba estadística, llamada Cociente C, se ha propuesto y permite evaluar asociaciones entre una gran cantidad de SNPs y un rasgo categórico, utilizando un estadística de prueba basado en los valores máximos de variables aleatorias χ^2. Esta nueva metodología ha demostrado en datos reales, encontrar un número mayor de polimorfismos asociados con el fenotipo y poseer mayor potencia estadística en comparación con las propuestas clásicas de corrección por pruebas múltiples.

Palabras clave: estudios de asociación del genoma completo, problemas de comparaciones múltiples, datos categóricos, error tipo I, potencia estadística.

List of tables

List of figures

Contents

1. Introduction

Multiple hypothesis testing refers to a method of testing in which more than one hypothesis is tested on the same data set simultaneously. This method of testing is not recently introduced, and the basis for such procedures was established since the forties and early fifties by Dunkan, Scheffé and Tukey. Their applicability due to technological advancement in recent years, specially in genomics, have become difficult and the limitations of the classical methods are more evident nowadays.

In several genomic data analysis contexts, a large number of statistical hypotheses are tested at the same time on the same dataset. For example, when genetic polymorphism is used as a marker and it wants to be associated with a phenotype. In the human genome, there are over one million polymorphic sites, they are called *Single Nucleotide Polymorphisms* or SNPs. A review found a 99.9% similarity in DNA sequence between two samples of people taken at random; the remaining 0.01% includes any difference or genome variation between individuals. These genetic variants or polymorphisms arise in mutations along different ancestors and have been used in studies to identify if any of them is associated with a particular phenotypic trait, such as a disease. However, over 90% of these genomic variations are neutral and are not able to explain a particular trait.

Studies that identify associations between SNPs and the presence of a trait (generally a disease or more generally an outcome) are called *Genome-wide association studies* or GWAS and compare whole genomes of two groups of participants using an epidemiological case-control design, without making assumptions about the genomic location of the possible causal variants. The DNA of individuals with the disease is compared to those free of it and the millions of genetic variations are read using arrays of SNPs. The allele frequencies or genotypes between these two groups are compared and whether a particular type of variant is more frequent in cases than in controls according to statistical tests, it is decided that the polymorphism is associated to the presence of the outcome.

The first GWAS was published in 2005 and it investigated the macular degeneration related to age. Although these studies have been used to evaluate associations between SNPs and traits mostly in humans, they serve also to answer the same question in other organisms, such as plants and animals. For example, association studies have helped to identify genetic variations associated with susceptibility of ornamental plants and crops to viruses or pests

as a strategy to improve production; in other cases, GWAS have been applied to identify genes or Single Nucleotide Polymorphism associated with traits of economic importance in animal productions, like genetic variations related to better quality of milk or meat.

The response variable can be a quantitative trait such as the weight of individuals or the amount of a certain hormone in blood, or it can be dichotomous as the presence or absence of the disease or trait under study. In the latter case, the results are summarized in a 2×2 contingency table with the presence of SNP as exposure and disease as the outcome. Traditionally, when a dichotomous variable is evaluated, the association between $k_i, i = 1, 2, \ldots, K$ SNPs and the outcome is tested using primarily a χ^2 test of homogeneity, followed, in some cases, by a logistic regression model in order to obtain a proper estimate of the association both raw and adjusted for covariates. Then, a correction of the $p-$values is performed for each one of the $H_1, H_2, \ldots, H_K$ hypotheses tested. The classical methods for those correction include Bonferroni´s, Holm´s, Hochberg´s proposal that control the *family-wise error rate* (FWER) and Benjamini & Hochberg´s one that control the *false discovery rate* (FDR). Other approaches have worked on the *false discovery proportion* and the construction of confidence bounds for this measure and recent advances have moved away of the classical view of the problem like the penalized regression, methods based on functional genomics data or on Bayesian statistics.

In GWAS the most difficult limitation to address is the limited statistical power achieved after making adjustment to the type I error for multiple comparisons and its impact on the number of false positives. Testing simultaneously and in an independent way a large number of markers potentially associated with an outcome may inflate the number of false positives and a very stringent correction can affect the probability of rejecting the null hypothesis when it is false or in GWAS, the probability of correctly detecting a genuine association (statistical power). To test simultaneously a great amount of SNPs, each one as a separate hypothesis, involves making adjustments to the significance level and indeed, increasing the sample size needed to reach an adequate power. That is not a realistic and possible situation in genomics because the entire or a great part of the genome of individuals must be sequenced and this is still expensive. Obtaining a solid criterion for declaring significant associations with high statistical power and without increasing the sample size is crucial to Genome-wide association studies and statistical solutions are needed.

Methods have been developed in order to address these limitations and they have reduced type I and type II errors. The proposed methods are mainly based on changing the type of test to establish the association and extending it for continuous traits. Some of them are very complex statistical methods and are difficult to use for non-statisticians who usually produce the data. Moreover, very few methods have focused on developing a new statistical test for categorical data, which is the most common form of measuring phenotypical traits

in humans and other organisms.

A new statistical test was proposed, called Quotient C test. It allows testing associations between thousands of SNPs and a categorical trait using an test statistic based on the maximum of χ^2 random variables. The results on real dataset show that the Quotient C represents a less stringent criterion and allows to declare associated with the outcome a large amount of Single Nucleotide Polymorphism compared to the classical methods used to correct for multiple testing, keeping the probability of incorrectly rejecting a true null hypothesis (type I error) equal or less than α. The proposed methodology has a lower type II error rate and indeed a better statistical power than Bonferroni´s, Holm´s, Hochberg´s and Benjamini & Hochberg´s methods. To improve the statistical power without increasing the simple size represents a clear advantage of the Quotient C compared to the other proposals and allows to minimizing the amount of false negatives.

2. Description and formulation of the problem

2.1. Background

Multiple hypothesis testing refers to the fact that more than one hypothesis is tested simultaneously on the same data set. Although this issue is not new and the basis for such procedures raised since the forties and first half of the fifties by Dunkan [1, 2], Scheffé and Tukey [3, 4], their applicability due to technological advances in recent years, mainly in the human genetics becomes difficult and the limitations of the classical methods are more and more evident nowadays.

With the advancement in knowledge regarding DNA since the early fifties and the human genetics from the genome project since 1990, it has been established that although this molecule is a constant in all humans, there are over one million sites that vary genetically among individuals [5], they are called *Single Nucleotide Polymorphisms* or SNPs. A review found a 99.9% similarity in DNA sequence between two samples of people taken at random of the same population; the remaining 0.01% include any difference or genome variation between individuals [6]. These genetic variants or polymorphisms, arise in mutations from different ancestors [7] and have been used in different studies to identify if any of them is associated with a particular trait [8], such as a disease. However, over 90% of these genomic variations are neutral and are not able to explain a particular trait [9].

A substantial number of SNPs have been identified across the genome through the HapMap project and these have been recorded in databases accessible for consultation by the scientific community like the dbSNP database, GWASdb or the National Human Genome Research Institute [10, 11].

Studies that identify associations between polymorphisms and the presence of a trait (generally a disease) are called *Genome-wide association studies* or GWAS and generally they compare entire genomes of two groups of participants using an epidemiological case-control design, without making assumptions about the genomic location of the possible causal variants. A case-control study is designed to determine if an exposure is associated with an outcome (i.e., a disease, a trait or any other condition of interest) [12]. This kind of studies

represent a sample strategy in which the most important characteristic is that the studied population is made up of at least two groups: one of them called case (a group known to have the outcome) and the other called the control group (a group known to be free of the outcome) [12, 13]. By definition, a case-control study is always retrospective because it starts with the outcome of interest, then traces back to investigate multiple exposures, so, it looks back in time to measure which subjects in each group were exposed, comparing its frequency in the case group to the control group [12, 13, 14, 15].

In GWAS, the DNA of people with the disease or the trait is compared to those free of it and the millions of genetic variations are read using arrays of SNPs. In general, the allele frequencies or genotypes between these two groups are compared and whether a particular type of variant is more frequent in cases than in controls according to statistical tests, it is decided that the SNP is associated to the presence of the disease or trait [16]. As the genome does not change in a subject since its birth, the SNPs are always considered as a retrospective exposure, so the case-control design is appropriated.

The response variable can be a quantitative trait such like the weight of individuals or the amount of a certain hormone in blood, or it can be dichotomous like the presence or absence of the disease or trait under study [17]. In these latter cases, the results are summarized in a 2×2 contingency table with the presence of SNP as exposure and the disease as the outcome. For example, to assess the association between the presence of the 4889A/G polymorphism of CYP1A1 gene and the presence of lung cancer [18], the results would be summarized as shown in Table (**2-1**).

Table **2-1**.: A hypothetic 2×2 contingency table to evaluate the association between a particular SNP (4889A/G) and the presence of lung cancer.

	Lung cancer		
Presence of the 4889A / G polymorphism	Yes	No	
Yes	a	b	$a+b$
No	c	d	$c+d$
	$a+c$	$b+d$	n

The association between a potential exposure and the outcome is evaluated using primarily a χ^2 test of homogeneity, followed in some cases by a logistic regression model in order to obtain a proper estimate of the association both raw and adjusted for covariates using *odd ratios* (OR) [19]. Thus, for each of the SNPs investigated, a table like the one shown above is constructed and then, the association is evaluated using an individual hypothesis for each

polymorphism; at the end, the level of significance for multiple testing is corrected [8]. Thus, a GWAS results in multiple contrast of K hypotheses between dichotomous categorical variables.

The first GWAS was published in 2005 and it investigated the macular degeneration related to age [20]. Although these studies have been used to asses associations between SNPs and traits mostly in humans, they also have served to establish genotype-phenotype associations in other organisms, like plants and animals. For example, association studies have helped to identify genetic variations associated with susceptibility of ornamental plants and crops to viruses or pests as a strategy to improve production [22, 23, 24]; in other cases, Genome-wide association studies have been applied to identify genes or SNPs associated with traits of economic importance in animal productions, like genetic variations related with better quality of milk or meat [25, 26, 27, 28].

The markers in the public domain available to support studies like GWAS exceed 1.4 million [5] and are available in databases to be consulted. In the Single Nucleotide Polymorphism Database (dbSNP), a free public archive and general catalog of identified genome variations developed and hosted by the National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI), there is information about 9 million SNPs (within and across different species), including most of the estimated 11 million SNPs with rare allele frequency less than 1% or greater that are estimated to exist in the human genome [29, 30]. Other databases are the SeattleSNPs PGA (focused on identifying, genotyping and modeling the associations between Single Nucleotide Polymorphisms in candidate genes and pathways that underlie inflammatory responses in humans) [31] and EVOLution of TREEs as drivers of terrestrial biodiversity (information about sets of DNA samples for ecological research) [32], among others.

To test simultaneously a large number of markers, each one as a separate hypothesis involves making adjustments to the significance level and increasing the sample size needed to correctly detect a genuine statistical association. Nevertheless, sample size increase is very costly because the entire genome of individuals must be sequenced and it still being very expensive, therefore statistical solutions are needed.

2.2. The statistical problem in GWAS

When a hypothesis is tested, the approach is as follows: let X be a random variable with probability mass function or probability density function $f(x|\theta)$, where θ is an unknown parameter but it is known to belong to a parameter space Θ, namely $\theta \in \Theta$ with $\Theta \subset R$. We want to test the null hypothesis $H_0 = \theta \subset \Theta_0$ against the alternative hypothesis $H_1 = \theta \subset \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$, based on a statistical test $T(X)$ [33, 34]. To pose

the hypothesis, a rejection region Γ is definided and Γ corresponds to a subset of the sample space that leads to the rejection of H_0, such that if the observed statistical $T(X) = t \in \Gamma$, the null hypothesis is rejected, otherwise data support H_1. Γ is chosen in such way that the probability of making a type I error, i.e rejecting a true null hypothesis is such small as is possible; this level denoted as α is known as the significance level of the test [34].

When more than one hypothesis is tested on the same dataset, the problem arises when all hypotheses are evaluated together at a given level α, so regardless of the multiplicity of the problem, the probability of making a type I error increases when the number of hypotheses increases. Thus, when contrasting simultaneously K null hypotheses, it is very likely that if all of them are true (not associated with the outcome) and some of them are simply rejected by chance. If K is large enough, it will always prove a rejected hypothesis although all are true [35].

In GWAS the most difficult limitation to address is the limited statistical power achieved after making adjustment to the type I error for multiple comparisons and their impact on the number of false positives and negatives [36, 37, 38, 39]. Testing simultaneously a large number of markers potentially associated with the outcome may inflate the number of false positives and to affect the probability of rejecting the null hypothesis when it is false or in GWAS, the probability of correctly detecting a genuine association genotype-phenotype (statistical power) [40].

In this kind of studies, usually millions of comparisons are performed between categorical data on the same samples of patients. Under this assumption, at a given level α, a number K of markers being tested and assuming that all hypotheses are independent, it is possible to calculate the probability P of making a type I error using the following formula $P = 1 - (1 - \alpha)^K$ [40, 46]. In this way, with a type I error of 0.05 and five SNPs tested simultaneously, there is a 22.61% of probability that one of them is associated significantly by mistake with the disease, probability that rises to 51.23% with 14 SNPs and to 90.05% with 45 ones.

In order to address such limitations generated in multiple comparison tests, techniques have been developed which have inevitably led to two undesirable outcomes: to increased levels of false positives by a weak correction or to decrease statistical power due to a very strong correction [40].

Consider the following mathematical structure of the multiple testing problems in GWAS: let $\mathcal{H} = (H_1^0, H_2^0, \ldots, H_K^0)$ be the collection of null hypotheses that we would like to reject for $k_i, i = 1, 2, \ldots, K$ Single Nucleotide Polymorphisms studied. An unknown number k_0 of these hypotheses is true whereas the other $k_1 = K - k_0$ is false. Now, let $\mathcal{T} \subseteq \mathcal{H}$ be the

collection of true hypotheses and $\mathcal{F} = \mathcal{H} \setminus \mathcal{T}$ the collection of false hypotheses. The objective of multiple comparisons procedures is to choose a collection $\mathcal{R} \subseteq \mathcal{H}$ of hypothesis to reject. If the $p-$values $p_1, p_2, \ldots, p_K$ for each of the hypotheses H_k^0 are calculated, one possibility for $\mathcal{R}$ is the collection $\mathcal{R} = \{H_k : p_k \leq T\}$, where all the H_k^0 are rejected (also called as a discovery) if its $p_k \leq T$; in this way, multiple comparison procedures can be simplified as the choice of a T [41]. Ideally $\mathcal{R} = \mathcal{F}$, however, this does not always happen and represents errors committed on rejected hypotheses. The number of errors and non-errors committed in a study with multiple testing of hypotheses, using a frequentist setting, are summarized in a table as shown in (**2-2**) [41, 44].

Table **2-2**.: Number of errors and non-errors committed in a study with multiple testing procedures.

	Null hypothesis		
Number of	Number not rejected	Number rejected	Total
True null hypotheses	$U = k_0 - V$	V	k_0
Non-true null hypotheses	$M = k_1 - U$	S	k_1
	$K - R$	R	K

For example, rejecting a true null hypothesis (type I error) or committing a false positive (or false discovery) corresponds to the V hypothesis of the set $\mathcal{R} \cap \mathcal{T}$; on the other hand, a false negative or type II error occurs when a null hypothesis that is false is not rejected, ie the $k_1 - U$ hypothesis of the set $\mathcal{F} \setminus \mathcal{R}$, where U are the true null hypotheses that are not rejected [41, 44].

Using this notation, the specific number K is known, the number k_0 and $k_1 = K - k_0$ of true and false null hypotheses are unknown parameters, while $R = \#\mathcal{R}$ is an observable random variable and S, M, U and V are unobservable random variables [47].

The particularity of each techniques of correction by multiple comparisons depends on the proposed T and the classical methods include the methods of Bonferroni, Holm, Hochberg and Benjamini & Hochberg. The most commonly used technique is the Bonferroni´s correction [8, 41], which involves dividing the significance level α by the number of hypotheses tested and the corrected $p-$values are used to compare the results [51], thus $T = \alpha/K$. In this method, $\alpha' = \frac{\alpha}{K}$ is computed and each individual hypothesis H_i with $p_i \leq \alpha'$ is rejected, so the total type I error is divided equally among all the hypotheses tested [52]. Reducing the significance level could eventually find new associations (increasing statistical power), but it is risky because false positives could be among the associated SNPs. In a similar way that Bonferroni´s method (using a single-step adjustment), Sidák proposed $T = 1-(1-\alpha)^{1/K}$ [53].

Other proposes include a sequential variation using the ordered $p-$values $p_{(1)}, \ldots, p_{(K)}$. In those methods, each $p-$value $p_{(i)}$ is compared with its critical value (T). So, Holm proposed $T = \alpha/(K-i)$ [54], Hochberg a $T = \alpha/(K-i+1)$ [55] and Benjamini and Hochberg a $T = i\alpha/K$ [56].

Bonferroni´s, Holm´s, and Hochberg´s proposal are considered as *familywise error rate* (FWER) controlling methods, that is they calculate the probability that $\mathcal{R}$ contains any error or $P(V > 0)$ [41, 54, 55]. Moreover, the procedure of Benjamini and Hochberg controls the *false discovery rate* (FDR) or the expected proportion of errors among $\mathcal{R}$) $(E(Q))$ [56].

2.2.1. Assumptions of multiple testing procedures

In multiple testing problems, an important assumption must be done about the dependence structure of the variables or the $p-$values. High-dimensional studies like GWAS rarely involve the analysis of independent variables, rather, many related variables are analyzed simultaneously; however, some of the classical methods for performing multiple testing rely on independence, or some form of weak dependence among the data corresponding to the variables being tested. Leek and Storey [57] have developed an approach for addressing arbitrarily strong multiple testing dependence in two cases: at the level of the original data collected (population-level) and before test statistics or $p-$values are been calculated (estimation-level); these cases are types of multivariate dependence between vectors. The authors assume that K related hypotheses tests are simultaneously performed, each one based on an $n-$vector sampled from a common probability space R^n and the data corresponding to each hypothesis test is the vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$ for $i = 1, 2, \ldots, K$. The dataset is able to be arranged into a $K \times n$ matrix $\mathbf{X}$ where the $i-$th row is composed of $\mathbf{x}_i$, the outcome for each one of the hypothesis is arranged in a vector $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ and the goal of the procedure is to perform a hypothesis test on $E[\mathbf{x}_i|\mathbf{Y}]$. With respect to $E[\mathbf{x}_i|\mathbf{Y}]$, it is able to be modeled using linear models, nonparametric smoothers, longitudinal models, among others, so this expression can be written as $E[\mathbf{x}_i|\mathbf{Y}] = \mathbf{b}_i\mathbf{S}(\mathbf{Y})$, where $\mathbf{b}_i$ is a $1 \times d-$vector and $\mathbf{S}(\mathbf{Y})$ is a $d \times n$ matrix of basics functions evaluated at $\mathbf{Y}d < n$. The vector $\mathbf{Y}$ is made up of any type of random variables and in GWAS, it is made up of dichotomous variables. Now, let be $\mathbf{E}$ the $K \times n$ matrix, so the model can be written as $\mathbf{X} = \hat{\mathbf{B}}\mathbf{S}(\mathbf{Y}) + \mathbf{E}$.

Definition 2.2.1 *Population-level multiple testing dependence. The population-level multiple testing dependence exist when*

$$Pr(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K|\, \boldsymbol{Y}) \neq Pr(\mathbf{x}_1|\mathbf{Y}) \times Pr(\mathbf{x}_2|\mathbf{Y}) \times, \ldots, \times Pr(\mathbf{x}_K|\mathbf{Y}). \tag{2-1}$$

Definition 2.2.2 *Estimation-level multiple testing dependence. The estimation-level multiple testing dependence exists when*

$$Pr(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K|\hat{\mathbf{B}}, \mathbf{S}(\mathbf{Y}) \neq Pr(\mathbf{x}_1|\hat{\mathbf{B}}, \mathbf{S}(\mathbf{Y})) \times, \ldots, \times Pr(\mathbf{x}_K|\hat{\mathbf{B}}, \mathbf{S}(\mathbf{Y})). \tag{2-2}$$

From (2-1), the dependence at the population level is therefore any probabilistic dependence among the $\mathbf{x}_i$, after conditioning on $\mathbf{Y}$ and this is equivalent to the existence of dependence across the rows of $\mathbf{E}$. Meanwhile, from (2-2), the estimation-level dependence is equivalent to dependence among the rows of the residual matrix $\mathbf{R} = \mathbf{X} - \hat{\mathbf{B}}\mathbf{S}(\mathbf{Y})$. When population-level multiple testing dependence exists, then this will lead to estimation-level multiple testing dependence [57].

Typically, multiple testing dependence has been defined in terms of $p-$values or test statistics resulting from multiple tests. Let is assume the collection $S_1, S_2, \ldots, S_K$ of test statistics, one for each hypothesis tested and its corresponding unadjusted $p-$values $p_1, p_2, \ldots, p_K$. The assumption over the $p-$vales can be in two mainly ways: (i) non assumption at all and (ii) some assumption on the dependence structure of the $p-$values known as *dependence through stochastic ordering* or PDS [41]. With respect to the last one, it involves only the $p-$values of the true hypotheses, denoted as $q_1, q_2, \ldots, q_{K_0}$. The PDS comes in two ways: a weaker and a stronger one [58, 59]. The weaker PDS condition says that

$$E[f(q_1, q_2, \ldots, q_{K0})|q_i = u] \tag{2-3}$$

is nondecreasing in u for every i and for every f, a coordinate-wise nondecreasing function. The stronger PDS condition is similar to (2-3) but involves p_i instead of q_i, that is

$$E[f(p_1, p_2, \ldots, p_K)|q_i = u] \tag{2-4}$$

In (2-4), there are nondecreasing function f of $p-$values for all the hypotheses, not only for the true ones [58, 59].

2.2.2. Recent advances and challenges

As a step further to the classical methods, mainly FRD-controlling ones, other approaches have worked on the *false discovery proportion* or FDP denoted as Q and the construction of confidence bounds for this measure. The FDP is the proportion of false discoveries among total rejections. Formally it is defined as:

$$Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2-5}$$

Storey´s approach [60] considers the set $\mathcal{R}$ of rejected hypotheses of the form $\mathcal{R} = \{H_k : p_k \leq t\}$ where t is a constant. In $\mathcal{R}$, Storey proposes that the expected number of true hypothesis to be rejected is $k_0 t$ and an estimation of k_0, denoted as $\hat{k_0}$, is given by $\hat{k_0} = \frac{\#\{p_k > \lambda\}+1}{1-\lambda}$, where $0 \leq \lambda < 1$. The value of λ usually is proposed as 0.5 although $\lambda = \alpha$ is an usual choice. So,

the estimation of Q, denoted as $\hat{Q}$ is given by $\hat{Q} = \frac{\hat{k_0}t}{\#R}$[60].

Efron proposed a view of the FDP and the $FDR = E[Q]$ called *local FDR* [61]. The Storey method allows obtain conclusions only over the set $\mathcal{R}$ and not over any subset of $\mathcal{R}$ or over independently rejected hypothesis. The Efron´s approach solves this limitation but it comes with additional assumptions. The proposal assumes the test statistics $S_1, S_2, \ldots, S_K$ were randomly drawn from a mixture distribution with density $f(t) = \pi_0 f_0(t) + (1 - \pi_0) f_1(t)$, where π_0 is the probability that a hypothesis is true, $f_0(t)$ is the density of the test statistics of true hypotheses and $f_1(t)$ is the density of test statistics of false hypotheses. The shape of f_1 and π_0 are unknown and both must be estimated from the data meanwhile f_0 is considered as known. So, a similar version of FDR, the $Pr(H_i \text{ is true } | S_i \geq t)$ is calculated as $Pr(H_i \text{ is true } | S_i \geq t) = \frac{\pi_0 f_0(t)}{f(t)}$ [61].

A confidence estimation of the FDP is sometimes more informative than a point estimation, mostly in cases of variability. Two works are well-known to address this situation: Goeman and Solaris [62], developed an interval estimation of FDP under the PDS assumption and Meinshausen developed confidence intervals using a permutation-based method [63].

Modern proposals have moved away of the classical view of the multiple hypotheses testing problem. The penalized regression has been used to predict disease risk based on a genome, e.g. on a large number of single-nucleotide polymorphisms [64]. This methodology, although it has a different purpose (to predict the probability of a outcome based on the presence of SNPs), has incorporated techniques to select the SNPs to include in the model as *least absolute shrinkage and selection operator* or LASSO [65], which estimates a vector regression coefficients associated with the SNP effects by maximizing a penalized log-likelihood function and at the end, the not selected SNPs in the final model have regression coefficients set to zero. Other techniques used for a similar purpose are the *smoothly clipped absolute deviation* (SCAD) [66] and the *truncated* L_1-*penalty* or (TLP) [67].

Kichaev et al. [68] presented a new proposal called *functionally informed novel discovery of risk loci* or FINDOR based on functional genomic data. This method uses polygenic modeling to weight SNPs according to how well they tag functional categories that are enriched for heritability and it can be summarized in two steps. First, a function that predicts the χ^2 statistic is estimated for each SNP using a comprehensive assortment of functional annotations which includes coding, conserved and regulatory annotation. This random variable is predicted using a stratified LD score regression that includes as input a measure of confounding biases, the effect size on per-SNP heritability of annotation and an LD score which indicates the degree to which each SNP tags the annotation. The second step consists in stratifying the SNPs based on their expected χ^2 values. In simulated data, the methodology has shown to improve in a 9%-38% the number of independent associations genotype and

indeed, to improve the statistical power [68].

Bayesian statistic has also been applied to the selection of SNPs associated with clinical or biological traits. The Bayes´s theorem is used to combine prior beliefs of marker effects, which are expressed in terms of prior distributions, with information from data for inference [69]. In this framework, the authors propose a regression model for the SNP joint effect on the response and to specify a prior for the regression coefficients. The best multiple SNP model is determined evaluating the posterior probability [70, 71, 72, 73, 74]. As well as in the classical proposals, the number of models evaluated is usually adjusted for the multiple testing [74] and the high-dimension is still a limitation of those methods. Some of them have been development only for $K = 2$ [75] or they require a heavy first step selection to reduce the number of SNPs [76].

The most obvious technical limitation of GWAS is the high cost and effort required to genotype hundreds of thousands of SNPs per individual; because of this, the most of the conducted studies have had modest sample sizes with a consequent reduction in the statistical power of the tests used. These small studies are prone to random error and often offer large confidence intervals to draw meaningful conclusions [77]. Experience in other clinical settings suggests that they tend to produce more favorable results compared to studies with large sample sizes [78]. To increase these sizes, meta-analyzes have been developed and they integrate structured and systematic GWAS results over a given SNP and disease. Although this practice certainly increases the size of the study, a meta-analysis combines the results from multiple studies in an effort to increase power (over individual studies) and in Genome-wide association studies, this has a important implication on the results because it increases the diversity of genetic variants in the control group compared to cases [79] due to multiple genetic origins of the participants.

GWAS have had great success in identifying genes associated with the development of monogenic disorders or Mendelian diseases (where a single gene is responsible for the disease), but in polygenic diseases, their success has been limited because the associated genes explain only a small portion of the risk [80]. This lack of success has been partly explained by the low heritability of polymorphisms, imprecise definition of phenotypes and the most important fact: the low statistical power of the test of hypothesis [81].

Minimizing false positives generated through multiple hypothesis testing in Genome-wide association studies is essential for obtaining a solid criterion for declaring significant associations with greater statistical power without increasing the sample size and without ignoring the high-dimensionality [81, 82]. Today, although the developed techniques have made an approach to the solution of these problems, the limitations hinder its use as a true practice, sacrificing the validity of the results. Given this gap in scientific knowledge, a statistical

methodology is needed to keep the probability of correctly detecting a true association and reduce false positive rate in multiple hypothesis testing between categorical variables with application to whole genome association studies and other areas of knowledge where issues of a similar nature arise.

3. Objectives

3.1. General objective

To develop a statistical methodology for estimating associations between categorical data in multiple contrasts of hypothesis conservative of statistical power and false positive rate, with application to studies of genome-wide association.

3.2. Specific objectives

- To develop a statistical methodology to determine associations between categorical variables in multiple hypothesis testing.
- To evaluate the usefulness of the proposed technique by estimating associations between SNPs and the development of a particular complex disease, using a real database.
- To determine the statistical power generated with the proposed methodology and compare these parameters to those obtained by conventional statistical techniques used in solving this problem.

4. Proposed methodology

4.1. Overview of the methodology

The methodology was proposed in two stages:

1. Theoretical phase. A new test statistic was proposed based on the maximum of the statistical test used to prove the hypotheses:

   H_i^0 : *The ith polymorphism is not associated with the outcome.*
   H_i^1 : *The ith polymorphism is associated with the outcome.*

   For simplicity, we denoted $U_i(k_i) = X_i$ as the value of the statistical test used to prove the previous hypotheses for $i = 1, 2, \ldots, K$ polymorphism. The necessary steps are described below:

   a) The probability density function ($p.d.f$) of $X_{(K)} = \max(X_1, X_2, \ldots, X_K)$ denoted as $g_k(x_k)$ was found.

   b) The parameter r^* of the function $g_k(x_k)$ was estimated.

   c) The quotient between the random variable $U_i(k_i)$ and the density function of the maximum evaluated in $X_i = x_i$ was proposed as statistical test to choose the SNPs that are associated with the outcome. This statistical test was called Quotient C and formally, it is defined as:

   $$C = \frac{U_i(k_i)}{g_k(x_i)}, g_k(x_i) > 0. \tag{4-1}$$

   d) The $p.d.f$ of the Quotient C was found. First, the joint density function between $X := U_i(k_i) \sim \chi_1^2$ and $Y := g_{(k)}(x_i)$ is found and then, the joint density function between X and C denoted as $u_{XC}(x, c)$. The probability density function of C, denoted as $f_C(c)$ was estimated (by transformation techniques of random variables), integrating $u_{XC}(x, c)$ with respect to x, that is $f_C(c) = \int_{-\infty}^{\infty} u_{XC}(x, c)dx$.

2. Application phase. The proposed methodology has been applied to real data in comparison to the classical methods of multiple testing, following the next steps:

a) The classical methods of correction by multiple comparison tests as well the proposed methodology were applied to two real databases and the results were compared.

b) The statistical power of Quotient C as well as the other methods was calculated using the approach proposed by Dudoit, Boldrick and Shaffer [44].

c) A simulation study was performed in order to establish the repeatably of the results. In this case, three methods were compared: Bonferroni, Benjamini and Hochberg and Quotient C. Three scenarios were considered: i. SNPs detected by all the methods. ii. SNPs detected only by Benjamini and Hochberg´s proposal and iii. SNPs detected only by the Quotient C. In each scenario, a sample of 30% of polymorphism was taken. The simulation included 1000 repetitions in which the three methods were applied. The objective was to count the number of SNPs that each method found as associated with the outcome and summarize how many of them was detected by the other methods according to the scenario.

Figure (**4-1**) shows the schematic representation of the proposed methodology. +

4.2. Data sets

4.2.1. Single Nucleotide Polymorphism associated with hardness of meat in sheep

This dataset was supplied by Ariza et al. [48] and has the genetic variants studied as associated factors with the hardness of meat in sheep. In 130 sheep of the breed *Camuro* or *ovino de pelo criollo colombiano*, sacrificed under controlled conditions, the muscular hardness of *Longgisimus dorsi* muscle was studied as response variable. The hardness of the meat was considered as hard if the shear force value, using a Warner-Brattzler machine, was equal or higher than 17.118 u that is ($Y = 1$); otherwise, the meat was considered tender ($Y = 0$).

The genotyping of each animal was done with the chip OvineSNP50 BeadChip [49], which identified 30.776 SNPs in the entire ovine genome with an average gap size of 50.9 kb and a median gap size of 42.5 kb; this level of coverage provides excellent SNP density to power robust genome-association studies in sheep [49].

4.2.2. SNPs from HLA-DRB1 associated with multiple sclerosis in an admixed Colombian population

This dataset was supplied by Toro et al. [50] and has the information about SNPs from the *major histocompatibility complex, class II, DR beta 1* or HLA-DRB1, studied as associated factors with multiple sclerosis (MS, clinically defined according to the 2005 McDonald criteria) in a group of patients from Bogotá, Colombia. Using a case-control design, with 100 subjects with MS and 200 controls (ratio case/control: 1:2), 13 SNPs were sequenced using the FlexiGene DNA kit (QIAGEN).

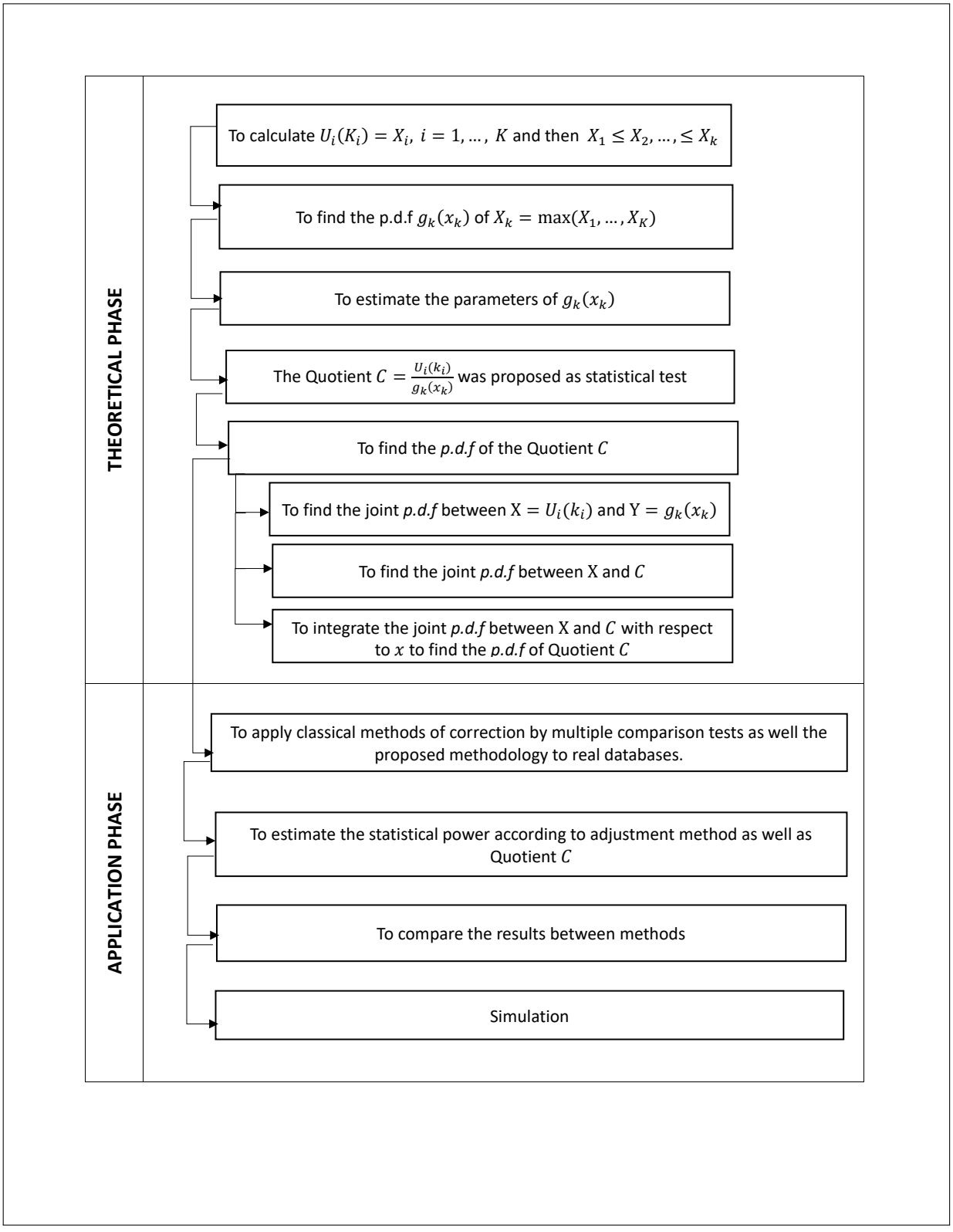

Figure **4-1**.: General diagram of the proposed methodology.

5. Results

5.1. χ^2 test in GWAS

Let us assume that the data have been collected without error in an epidemiological case-control design with $k_i, i = 1, 2, \ldots, K$ SNPs studied, all of them dichotomous and nominal (presence or absence of the rare allele in the DNA sequence of the individual). Let be n the sample size, $n_1 = \theta n$ the number of cases and $n_2 = (1 - \theta)n$ the number of controls, with θ as the proportion of cases in n.

Let j be the index of a subject in the sample and for the $j-$th subject, $Y = t$ denotes if he/she is a case or a control. So, $Y = 1$ is the $j-$th subject is a case or $Y = 0$ if the individual is a control. Usually, to evaluate the association between the ith genetic variation and the outcome, a 2×2 table is elaborated and the statistical test used to test the hypotheses H_i^0 : the ith SNP is not associated with $Y = t$, again H_i^1 : the ith SNP is associated with $Y = t$, is given by

$$U_i(k_i) = \frac{n(ad - bc)^2}{(a + b)(a + c)(c + d)(b + d)} \tag{5-1}$$

with a, b, c y d denoted as in the Table (**2-1**).

With K independent polymorphism and a true H_i^0 , $U_i(k_i) \sim \chi_1^2$ and $U_i(k_i) \in [0, \infty)$. The biggest values of $U_i(k_i)$ lead to reject the null hypothesis.

5.2. Probability density function of the maximum of χ^2 random variables

For simplicity, let is denote $X_i = U_i(k_i), i = 1, 2, \ldots, K$ and we can see that $X_1, X_2, \ldots, X_K$ are independent and identically distributed random variables (iid), each one $X_i \sim \chi_{(r)}^2$

with $r = 1$ degrees of freedom. Let $F(x)$ and $f(x)$ be the common cumulative distribution functions and probability density function, respectively, of these random variables.

Now, let $X_{(1)}, X_{(2)}, \ldots, X_{(K)}$ be the ordered random variables, where $X_{(1)} \leq X_{(2)} \leq , \ldots, \leq X_{(K)}$. Using this notation, $X_{(1)} = \min(X_1, X_2, \ldots, X_k)$ and $X_{(K)} = \max(X_1, X_2, \ldots, X_K)$. The probability density function of $X_{(K)}$ is given by the following theorem:

Theorem 1 *Density function of the maximum of order statistics [84]. Let $X_1, X_2, \ldots, X_K$ be independent identically distributed continuous random variables with common distribution function $F(x)$ and common density function $f(x)$. If $X_{(k)}$ denotes the maximum of $X_1, X_2, \ldots, X_K$, then the density function of $X_{(k)}$ is given by*

$$g_k(x_k) = K\left[F(x)\right]^{K-1} f(x). \tag{5-2}$$

As $X_1, X_2, \ldots, X_K$ are continuous and iid random variables, each one of them $\sim \chi^2$ and using (5-2), the maximum of those variables have the probability density function given by

$$\begin{aligned} g_k(x_k) &= K\left[\frac{\gamma(r/2, x/2)}{\Gamma(r/2)}\right]^{K-1} \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2} 1_{(0,\infty)}(x) \\ &= \frac{K\left[\gamma(r/2, x/2)\right]^{K-1}}{\Gamma(r/2)^{K-1}} \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2} 1_{(0,\infty)}(x) \\ &= \frac{K\left[\gamma(r^*/2, x/2)\right]^{K-1}}{2^{r^*/2}\Gamma(r^*/2)^K} x^{r^*/2-1} e^{-x/2} 1_{(0,\infty)}(x), \end{aligned} \tag{5-3}$$

where $\gamma(r/2, x/2) = \int_0^{x/2} t^{r/2-1} e^{-t} dt$ is an incomplete gamma function and $\Gamma(r/2) = \int_0^{\infty} t^{r/2-1} e^{-t} dt$ is a gamma function, which by definition is a constant. So, the density function given in (5-3) follows a distribution with r^* degrees of freedom.

5.3. Proposed statistical test

5.3.1. Proposed test statistic

When $k_i, i = 1, 2, \ldots, k, \ldots K$ polymorphisms are tested and the random variables X_i are considered as order statistics, the interest is choosing those X_i with the biggest values that lead to the rejection of the null hypotheses at α or the type I error. Considering the distribution of the maximum of the random variables $X_1, X_2, \ldots, X_K$ and

to evaluate the values of $X_i = x_i$ under this distribution, we estimate the probability that each one of the random variables belongs to these maximum.

The quotient between the random variable $U_i(k_i)$ and the density function of the maximum evaluated in $X_i = x_i$ was proposed as the statistical test to choose the SNPs that are associated with an outcome. This statistical test was called Quotient C, and it is defined by:

$$C = \frac{U_i(k_i)}{g_{k_i}(x_i)}, g_{k_i}(x_i) > 0.$$

Based on this quotient, we deduce the following:

- If $C_i \geq 1$, then the value of the statistic $X_i = U_i(k_i)$ is at least equal or greater than the value of $X_i = x_i$ based on the distribution of the maximum value of the random variables, $X_1, X_2, \ldots, X_K$, assuming that X_i is the maximum.
- If $C_i < 1$, then it implies that the value of the statistic $X_i = U_i(k_i)$ is smaller than the value of $X_i = x_i$ based on the distribution of the maximum of the random variables, assuming that X_i is the maximum.

As a consequence of applying the Quotient C, the SNPs that are associated with the presence of the outcome were those whose $C_i \geq 1$ and are statistically significant at any given α; this is equivalent to the collection $\mathcal{R} = \{H_k : p_k \in X_{(K)}\}$.

5.3.2. Probability density function of Quotient C

The Quotient C is a function of two independent random variables: $X := U_i(k_i) \sim \chi^2_1$ and $Y := g_{(k)}(x_i)$. The latter follows the distribution given in (5-3). In this way, the joint probability density function between X and Y is:

$$f_{XY}(x,y) = \begin{cases} \frac{1}{\Gamma(r/2)2^{\frac{r}{2}}} x^{\frac{r}{2}-1} e^{\frac{-x}{2}} \frac{K\left[\gamma(\frac{r^*}{2}, \frac{y}{2})\right]^{K-1}}{2^{\frac{r^*}{2}}\Gamma(r^*2)^K} y^{\frac{r^*}{2}-1} e^{\frac{-y}{2}} & \text{if } 0 \leq x, 0 \leq y \\ 0 & \text{, otherwise.} \end{cases}$$

As $r = 1$ and $\Gamma(1/2) = \sqrt{\pi}$, then

$$
\begin{aligned}
f_{XY}(x,y) &= \begin{cases} \frac{1}{\Gamma(1/2)2^{\frac{1}{2}}} x^{\frac{-1}{2}} e^{\frac{-1}{2}} \frac{K\left[\gamma(\frac{r^*}{2},\frac{y}{2})\right]^{K-1}}{2^{\frac{r^*}{2}}\Gamma(r^*2)^n} y^{\frac{r^*}{2}-1} e^{\frac{-y}{2}} & \text{if } 0 \le x, 0 \le y \\ 0 & \text{, otherwise.} \end{cases} \\
&= \begin{cases} \frac{K\left[\gamma(\frac{r^*}{2},\frac{y}{2})\right]^{K-1}}{\sqrt{2\pi}2^{r^*/2}\Gamma(r^*/2)^K} x^{-1/2} e^{-1/2} y^{\frac{r^*}{2}-1} e^{\frac{-y}{2}} & \text{if } 0 \le x, 0 \le y \\ 0 & \text{, otherwise.} \end{cases}
\end{aligned} \tag{5-4}
$$

To find the distribution of the proposed statistical test, first, the joint density between X and C was found and then, the marginal probability density function of Quotient C.

The joint probability density function between X and C was found by the method of transformation [83, 84, 85], which is based on the following proposition:

Proposition 5.3.1 *Let X be a random $k-$vector with a Lebesgue p.d.f. fx and let $Y = g(X)$, where g is a Borel function from $(R^k, \mathcal{B}^k)$ to $(R^k, \mathcal{B}^k)$. Let $A_1, \ldots, A_m$ be disjoint sets in $\mathcal{B}^k$ such that $R^k - (A_1 \cup \ldots \cup A_m)$ has Lebesgue measure 0 and g on A_j is one-to-one with a non-vanishing Jacobian, i.e., the determinant $Det(\partial g(x)/\partial x) \neq 0$ on $A_j, j = 1, \ldots, m$. Then Y has the following Lebesgue p.d.f.:*

$$f_Y(x) = \sum_{j=1}^{m} |Det(\partial g(x)/\partial x| fx(h_j(x)),$$

where h_j is the inverse function of g on $A_j, j = 1, \ldots, m$.

Now, let be $h(x,y) = C$ and we assign X a value $x \geq 0$, then $C = \frac{x}{Y}$ y $Y = h^{-1}(c) = \frac{x}{c}$. On the other hand, we have that $\frac{\partial}{\partial c} h^{-1}(c) = \frac{\partial}{\partial c}\frac{x}{c} = \frac{-x}{c^2}$.

Let´s denote $u(x,c)$ as the joint probability density function between X and C,which is given by

$$
u_{XC}(x,c) = \begin{cases} f\left[x, h^{-1}(c)\right] \left|\frac{\partial}{\partial c} h^{-1}(c)\right| & \text{if } 0 \le x, 0 \le \frac{x}{c} \\ 0 & \text{, otherwise.} \end{cases} \tag{5-5}
$$

Replacing $y = \frac{x}{c}$ in (5-4), $f\left[x, h^{-1}(c)\right]$ is found and applying (5-5) we have:

$$u_{XC}(x,c) = \begin{cases} \frac{K\left[\gamma(\frac{r^*}{2},\frac{x/c}{2})\right]^{K-1}}{\sqrt{2\pi}2^{r^*/2}\Gamma(r^*/2)^K}x^{-1/2}e^{-1/2}\left(\frac{x}{c}\right)^{\frac{r^*}{2}-1}e^{\frac{-(x/c)}{2}}\left|\frac{-x}{c^2}\right| & \text{if } 0\le x, 0\le y \\ 0 & \text{, otherwise.} \end{cases} \quad (5\text{-}6)$$

Multiplying powers with the same base and applying properties of the exponents in (5-6):

$$\begin{aligned} \frac{K\left[\gamma(\frac{r^*}{2},\frac{x/c}{2})\right]^{K-1}}{\sqrt{2\pi}2^{r^*/2}\Gamma(r^*/2)^K}x^{-1/2}e^{-1/2}\left(\frac{x}{c}\right)^{\frac{r^*}{2}-1}e^{\frac{-(x/c)}{2}}\frac{x}{c^2} &= \frac{K\left[\gamma(r^*/2,\frac{2x}{c})\right]^{K-1}}{\sqrt{2\pi}2^{r^*/2}\Gamma(r^*/2)^K c^{\frac{r^*}{2}-1+2}}x^{\frac{-1}{2}+1+\frac{r^*}{2}-1}e^{-1/2}e^{\frac{-2x}{c}} \\ &= \frac{K\left[\gamma(r^*/2,\frac{2x}{c})\right]^{K-1}}{\sqrt{2\pi}2^{r^*/2}\Gamma(r^*/2)^K c^{\frac{r^*}{2}+1}}x^{\frac{r^*+1}{2}-1}e^{-1/2}e^{\frac{-2x}{c}} \end{aligned}$$

Then

$$u_{XC}(x,c) = \begin{cases} \frac{K\left[\gamma(r^*/2,\frac{2x}{c})\right]^{K-1}}{\sqrt{2\pi}2^{r^*/2}\Gamma(r^*/2)^K c^{\frac{r^*}{2}+1}}x^{\frac{r^*+1}{2}-1}e^{-1/2}e^{\frac{-2x}{c}} & \text{if } 0\le x, c>0 \\ 0 & \text{, otherwise.} \end{cases}$$

By integrating the joint density function between X and C with respect to x, Quotient C's p.d.f is given by

$$\begin{aligned} f_C(c) &= \int_{-\infty}^{\infty} u_{XC}(x,c)dx \\ &= \int_0^{\infty} \frac{K\left[\gamma(r^*/2,\frac{2x}{c})\right]^{K-1}}{\sqrt{2\pi}2^{r^*/2}\Gamma(r^*/2)^K c^{\frac{r^*}{2}+1}}x^{\frac{r^*+1}{2}-1}e^{-1/2}e^{\frac{-2x}{c}}dx \end{aligned}$$

Now, let's consider the following arrangement:

$$\begin{aligned} &= \int_0^{\infty} \frac{Ke^{-1/2}\left[\gamma(r^*/2,\frac{2x}{c})\right]^{K-1}}{\sqrt{2\pi}2^{r^*/2}\Gamma(r^*/2)^K c^{\frac{r^*}{2}+1}}\frac{\Gamma(\frac{r^*+1}{2})}{(\frac{2}{c})^{\frac{r^*+1}{2}}}\frac{(\frac{2}{c})^{\frac{r^*+1}{2}}}{\Gamma(\frac{r^*+1}{2})}x^{\frac{r^*+1}{2}-1}e^{\frac{-2x}{c}}dx \\ &= \frac{Ke^{-1/2}\Gamma(\frac{r^*+1}{2})}{\sqrt{2\pi}2^{r+1}\Gamma(r^*/2)^K\sqrt{c}}\int_0^{\infty}\left[\gamma(r^*/2,\frac{2x}{c})\right]^{K-1}g\left(x;\frac{r^*+1}{2},\frac{2}{c}\right)dx \\ &= \frac{Ke^{-1/2}\Gamma(\frac{r^*+1}{2})}{\sqrt{c2\pi}2^{r^*+1}\Gamma(r^*/2)^K}E_X\left\{\left[\gamma(r^*/2,\frac{2X}{c})\right]^{K-1}\right\} \quad (5\text{-}7) \end{aligned}$$

Where $g(\cdot)$ is the probability density function of a random variable that follows a gamma distribution with parameters $\left(\frac{r^*+1}{2}, \frac{2}{c}\right)$ and $X \sim g(\frac{r^*+1}{2}, \frac{2}{c})$. In this way, the probability density function of Quotient C is given by the expression (5-7).

Intuitively, the expression of $f_C(c)$ has the following limitations when $K \to \infty$: (i) $E_X \to \infty$ so $f_C(c)$ can take too large values and (ii) $\Gamma(r^*/2)^K \to \infty$ so $f_C(c) \to 0$. Computationally speaking, the maximum and minimum value that a conventional computer accepts are $1.7976e^{308}$ and $2.225e^{-308}$ respectively; bigger values are considered as `Inf` and a lower one as 0. In this way, for the estimation of the $p-$values for the SNPs, the following adjustment were done: (i) when the computer assigned an infinite value, it was replaced by $1.7976e^{308}$ and (ii) when the computer assigned a value of 0, it was replaced by $2.225e^{-308}$.

Estimation of the probability density function of Quotient C

The graphs (**A-1**) to (**A-4**) show the plot function of $f_C(c)$ for different values of r^* and K. Is possible to see when this adjustment is done due to values of K bigger than 308 (as 1000 and 10000), the probability density function of the Quotient C is affected (Figure **A-2** to **A-4**). Although an analytic expression of the probability density function $f_C(c)$ was found, it is not possible to calculate it exactly or it is very difficult to calculate it with the degree of precision desired due the limitations previously exposed, so a *Monte Carlo integration* was performed. The Figure (**5-1**) shows an illustration of *Monte Carlo integration* of the probability density function of Quotient C for the two datasets used and $r^* = 1$.

5.3.3. Estimation of r^*

Maximum likelihood estimation of r^*

The probability density function of the maximum is given by (5-3) and let $\ell(r^*)$ be the likelihood function of $g_k(x_k)$. The part of this function that involves the parameter is the kernel and since the maximization of the likelihood function is done with respect to the parameter, the rest is irrelevant. Therefore

$$\ell(r^*) = \frac{[\gamma(r^*/2, x/2)]^{K-1}}{2^{r^*/2}\Gamma(r^*/2)^K} x^{r^*/2-1}.$$

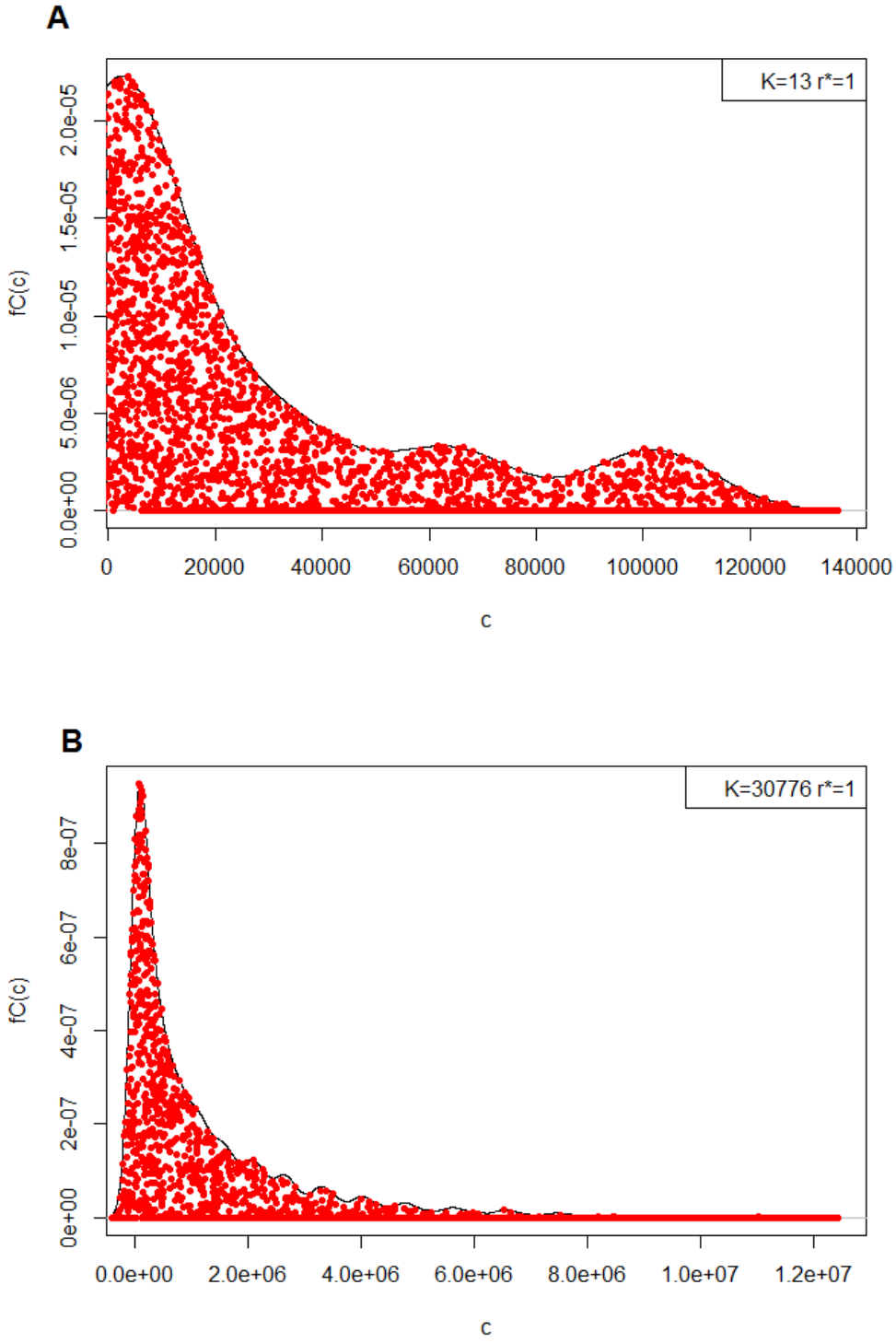

Figure **5-1**.: Illustration of the *Monte Carlo integration* when $r^* = 1$ for the dataset of Toro et al. [50] **(A)** and Ariza et al. [48] **(B)**.

Now, let be $L(r^*) = \ln[\ell(r^*)]$. Therefore

$$\begin{aligned} L(r^*) &= \ln\left[\frac{[\gamma(r^*/2, x/2)]^{K-1}}{2^{r^*/2}\Gamma(r^*/2)^K} x^{r^*/2-1}\right] \\ &= \ln\left[\gamma(r^*/2, x/2)^{K-1}\right] + \ln\left(x^{r^*/2-1}\right) - \ln\left(2^{r^*/2}\right) - \ln\left[\Gamma(r^*/2)^K\right] \\ &= (K-1)\ln[\gamma(r^*/2, x/2)] + \left(\frac{r^*}{2} - 1\right)\ln(x) - \frac{r^*}{2}\ln(2) - K\ln[\Gamma(r^*/2)] \quad (5\text{-}8) \end{aligned}$$

and

$$\begin{aligned} \frac{\partial L(r^*)}{\partial r^*} &= \frac{\partial}{\partial r^*}\left[(K-1)\ln[\gamma(r^*/2, x/2)] + \left(\frac{r^*}{2} - 1\right)\ln(x) - \frac{r^*}{2}\ln(2) - K\ln[\Gamma(r^*/2)]\right] \\ &= (K-1)\frac{\partial}{\partial r^*}\ln[\gamma(r^*/2, x/2)] + \frac{\ln(x)}{2} - \frac{\ln(2)}{2} - K\frac{\partial}{\partial r^*}\ln[\Gamma(r^*/2)]. \end{aligned}$$

How $\frac{\partial}{\partial z}[\ln\Gamma(z)] = \frac{\Gamma(z)'}{\Gamma(z)}$ (the Digamma function), then $K\frac{\partial}{\partial r^*}\ln[\Gamma(r^*/2)] = K\frac{\Gamma(r^*/2)'}{2\Gamma(r^*/2)}$, where $\Gamma(r^*/2)'$ is the ordinary derivative of $\Gamma(r^*/2)$ with respect to $r^*/2$. Now, how $\frac{\gamma(a,x)}{\Gamma(a)} = P(\chi^2|\nu)$ with $\nu = 2a$, $\chi^2 = 2x$ and

$$P(\chi^2|\nu) = \left[2^{\nu/2}\Gamma(\frac{\nu}{2})\right]^{-1}\int_0^{x^2}(t)^{\frac{\nu}{2}-1}e^{-\frac{t}{2}}dt_{1(0\leq\chi^2<\infty)}$$

we have that $\gamma(r^*/2, x/2) = P(\chi^2|r^*)\Gamma(r^*/2)$ with $\chi^2 = x$ [45]. So,

$$\begin{aligned} (K-1)\frac{\partial}{\partial r^*}\ln[\gamma(r^*/2, x/2)] &= (K-1)\frac{\partial}{\partial r^*}\ln\left[P(\chi^2|r^*)\Gamma(r^*/2)\right] \\ &= (K-1)\left[\frac{\partial}{\partial r^*}\ln P(\chi^2|r^*) + \frac{\partial}{\partial r^*}\ln\Gamma(r^*/2)\right] \\ &= (K-1)\left[\frac{\partial}{\partial r^*}\ln P(\chi^2|r^*) + \frac{\Gamma(r^*/2)'}{2\Gamma(r^*/2)}\right] \\ &= (K-1)\frac{\partial}{\partial r^*}\ln P(\chi^2|r^*) + (K-1)\frac{\Gamma(r^*/2)'}{2\Gamma(r^*/2)}. \end{aligned}$$

Now,

$$\begin{aligned} (K-1)\frac{\partial}{\partial r^*}\ln P(\chi^2|r^*) &= (K-1)\frac{\partial}{\partial r^*}\ln\left(\left[2^{r^*/2}\Gamma\left(\frac{r^*}{2}\right)\right]^{-1}\int_0^{x^2}(t)^{\frac{r^*}{2}-1}e^{-\frac{t}{2}}dt\right) \\ &= (K-1)\frac{\partial}{\partial r^*}\ln\left(\left[2^{r^*/2}\Gamma\left(\frac{r^*}{2}\right)\right]^{-1}\right) + (K-1)\frac{\partial}{\partial r^*}\ln\int_0^{x^2}(t)^{\frac{r^*}{2}-1}e^{-\frac{t}{2}}dt. \end{aligned}$$

Argument	Description	Value
f	The function to be optimized. The function is either minimized or maximized over its first argument.	Equation (5-8)
`lower`	The lower end point of the interval to be searched.	1
`upper`	The upper end point of the interval to be searched	100
`tol`	Tolerance, the desired accuracy.	0.0001220703

Table **5-1**.: Arguments used in the estimation of r^* through the function `optimize` in `R`.

We have that $(K-1)\frac{\partial}{\partial r^*}\ln\left(\left[2^{r^*/2}\Gamma\left(\frac{r^*}{2}\right)\right]^{-1}\right) = -(K-1)\frac{\Gamma(r^*/2)'}{4\Gamma(r^*/2)} - (K-1)\frac{\ln(2)}{2}$, so:

$$\frac{\partial L(r^*)}{\partial r^*} = -(K-1)\frac{\Gamma(r^*/2)'}{4\Gamma(r^*/2)} - (k-1)\frac{\ln(2)}{2} + (K-1)\frac{\partial}{\partial r^*}\ln\int_0^{x^2}(t)^{\frac{r^*}{2}-1}e^{-\frac{t}{2}}dt + \frac{\ln(x)}{2} - \frac{\ln(2)}{2} - K\frac{\Gamma(r^*/2)'}{2\Gamma(r^*/2)}. \quad (5\text{-}9)$$

As it is possible to see, the analytic solution to (5-9) is not simple, so the estimation of r^* was performed through computational methods.

Estimation of r^* through the function `optimize` in `R`

The function `optimize` of the package `stats` in `R - project` allows to find in an easy way a maximum (as well as minimum) of a function f with respect to its first argument, across $[a, b]$, an interval of possible values of the parameter to be estimated, where $a =$ `lower` (the minimum possible value to the parameter) and $b =$ `upper` (the maximum one); these are two arguments to be defined in `R`. The method used is a combination of *golden section search* and *successive parabolic interpolation* [42], and was designed for use with continuous functions [43].

The description of the arguments used are given in the Table (**5-1**). The $L(r^*)$ was used as the function f because it showed better behaviour when it was optimized. The appendix B shows the behaviour of f for different values of X_i and K.

5.4. Statistical power

The statistical power according to classical adjustment methods as well as quotient C was calculated using the approach proposed by Dudoit, Shaffer and Boldrick (see equation (5-11))[44]. They propose that the concept of statistical power can be generalized in different ways depending on single or multiple hypotheses. They define the statistical power in the following three ways [87]:

- The probability of rejecting at least one false null hypothesis, that is equal to $Pr(S \geq 1) = Pr(T \leq k_1 - 1)$.
- The average probability of rejecting the false null hypotheses or *average power* defined as $E(S)/k_1$, and
- The probability of rejecting all false null hypotheses, $Pr(S = m_1) = Pr(T = 0)$.

Also, the authors define the estimated statistical power, denoted as $\hat{\zeta}$, using the *False Discovery Rate* o FDR as:

$$\begin{aligned} \hat{\zeta} &= E\left(\frac{S}{R}|R > 0\right) Pr(R > 0) \\ &= Pr(R > 0) - FDR. \end{aligned} \tag{5-11}$$

The FDR, proposed by Benjamini and Hochberg [56] is defined as $FDR = E(Q)$, where Q is a measure of the number V of type I errors or the *false discovery proportion* (FDP). This last measure formally is definided as (2-5).

The *False Discovery Rate* is equivalent to:

$$FDR = E\left(\frac{V}{R}|R > 0\right) Pr(R > 0).$$

Also, $R \sim Bin(K, \pi)$ [41] where the likelihood estimation of π, denoted as $\hat{\pi}$ can be estimated as $\hat{\pi} = r/K$, where r is the number of null hypothesis rejected observed.

5.5. Application of the proposed methodology to the dataset of Single Nucleotide Polymorphism associated with hardness of meat in sheep.

5.5.1. General results

The study population was classified in 61 sheep with hard meat and 69 with tender one. The minor allele frequency ranged from 0.76% to 11.5% in the whole sample (mean = 5.37%) and from 0.76% to 10.9% in cases (mean = 5.07%) and 0.76% to 11.5% in controls (mean = 5.37%). Figure (**5-2**) shows the histogram of the raw (unadjusted) $p-$values of the χ^2 test for each one of the SNPs studied **(A)** and the plot of the sorted $p-$values **(B)**, respectively. On **(A)** it is possible to see an apparent non-uniform distribution of the $p-$values and a higher amount of values less than 0.01 as well as equal to 1. On the other hand, the plot of the sorted $p-$values shows that they follow

SNPs	Method					
	Non corrected	Bonferroni	Holm	Hochberg	B-H	Quotient C
Associated	4240	2606	2606	2606	2795	3787
Non associated	26536	28170	28170	28170	27981	26989
Total	30776	30776	30776	30776	30776	30776
Abbreviations: B-H, Benjamini and Hochberg´s method.						

Table **5-2**.: Number of SNPs associated and non associated in a raw and adjusted analysis as well by the Quotient C in the dataset of Ariza et al. [48].

approximately the dotted line, which means although there is evidence of a higher amount of SNPs with extreme values, in general, it is possible to expect a uniform profile of them. If an overwhelming majority of hypotheses were true, the histogram of the p-values would be approximately uniform and the plot of the sorted $p-$values approximately would follow the dotted line because p-values of true hypotheses follow a uniform distribution.

Table (**5-2**) shows the number of SNPs associated and non associated at $p \leq 0.05$ in an unadjusted analysis as well as according to classical multiple testing procedures and Quotient C.

Without any correction, the number of SNPs associated with hardness of meat in sheep was 4240, amount that decreases to 2606 when the procedure of Bonferroni, Holm or Hochberg is applied. These last methods detected 61.48% of the SNPs associated in a unadjusted analysis. The Benjamini and Hochberg´s correction identified a higher amount of Single Nucleotide Polymorphism associated with the outcome (2795) compared to FWER-based procedures, that represent a 7.22% of more polymorphism. On other hand, with respect to the Quotient C, this methodology identified as associated 3787 SNPs, a higher number compared to the classical methods (45.31% more than FWER-methods and 35.49% more than Benjamini and Hochberg´s one) and less than without correction. For illustrative purposes, Table (**5-3**) shows some of the SNPs detected as associated by Quotient C as test statistic and the comparison of its p-values with the values of the others methods; the SNPs were randomly selected and then sorted according its r^* values.

Figure (**5-3**) shows the Manhattan plot of the SNPs associated with the outcome along the entire genome of sheep in an unadjusted analysis **(A)** and by Bonferroni´s, Holm´s and Hochberg´s proposal **(B)**. It is possible to see a higher amount of polymorphism detected without correction compared to Bonferroni´s, Holm´s and Hochberg´s proposal. Also, as the polymorphism detected as associated with the FWER-controlling methods were the same, the Manhattan-plot were equal by the three procedures. On the other hand, the profile of SNPs associated with the hardness of meat was different when the Benjamini & Hochberg´s method was applied (Figure **5-4**, **(A)**), that implies not only the number of polymorphism associated is different when this method was used but also, the SNPs identified and their $p-$values. In Figure (**5-4**, **(B)**), the Manhattan-plot of Quotient C is presented and a clear difference not only in the number but also in the profile of the polymorphism is shown compared with the Manhattan-plot of the other proposals.

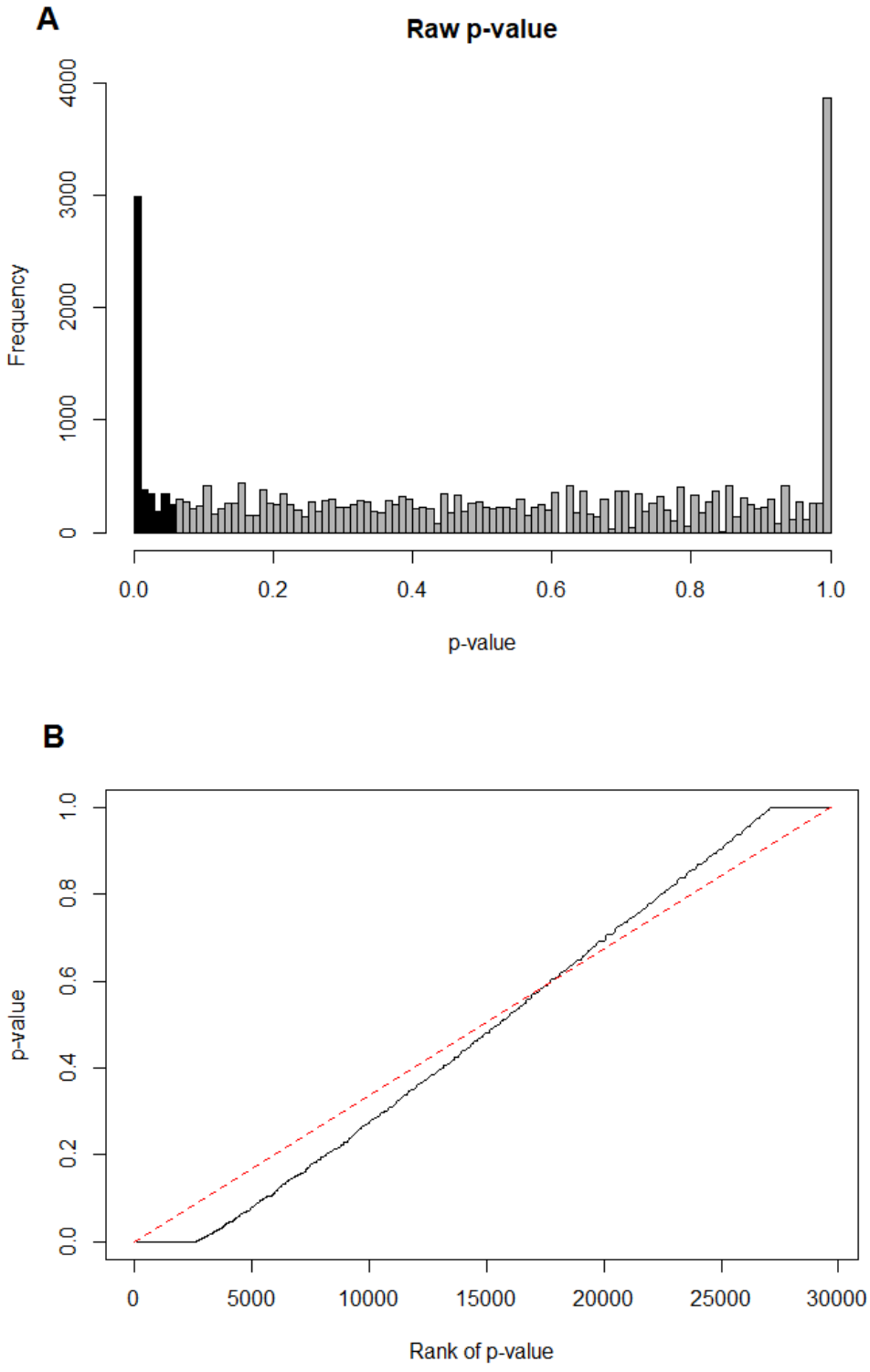

Figure **5-2**.: Histogram of the raw $p-$values **(A)** and profile of the raw $p-$values **(B)** of the dataset of Ariza et al. [48]. In **(A)**, the black bars represent SNPs with $p \leq 0.05$.

.

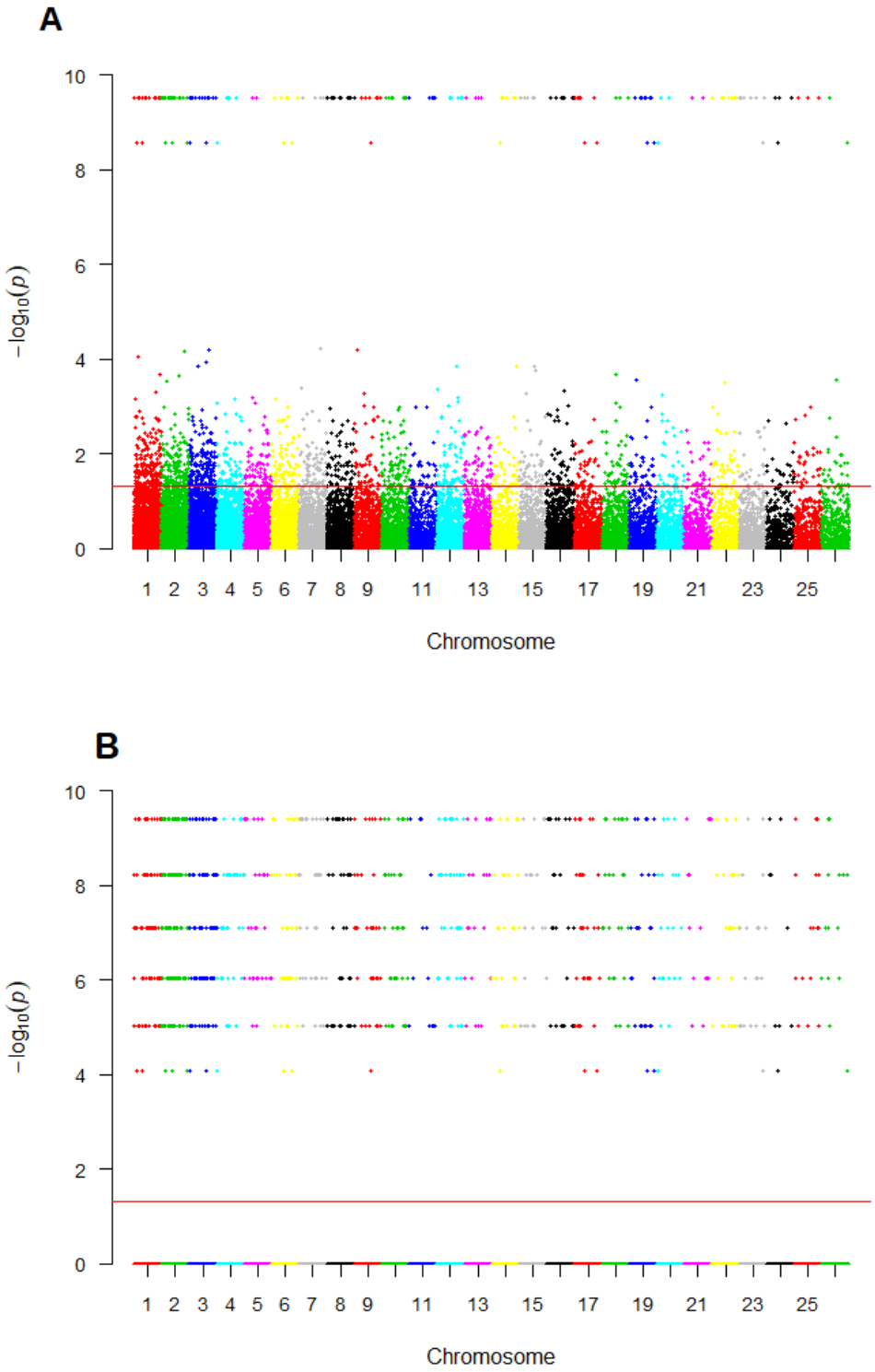

Figure **5-3**.: Manhattan plot of the SNPs associated with the hardness of meet in sheep in an unadjusted analysis **(A)** and FWER- based methods **(B)**. The red solid line represents $p = 0.05$.

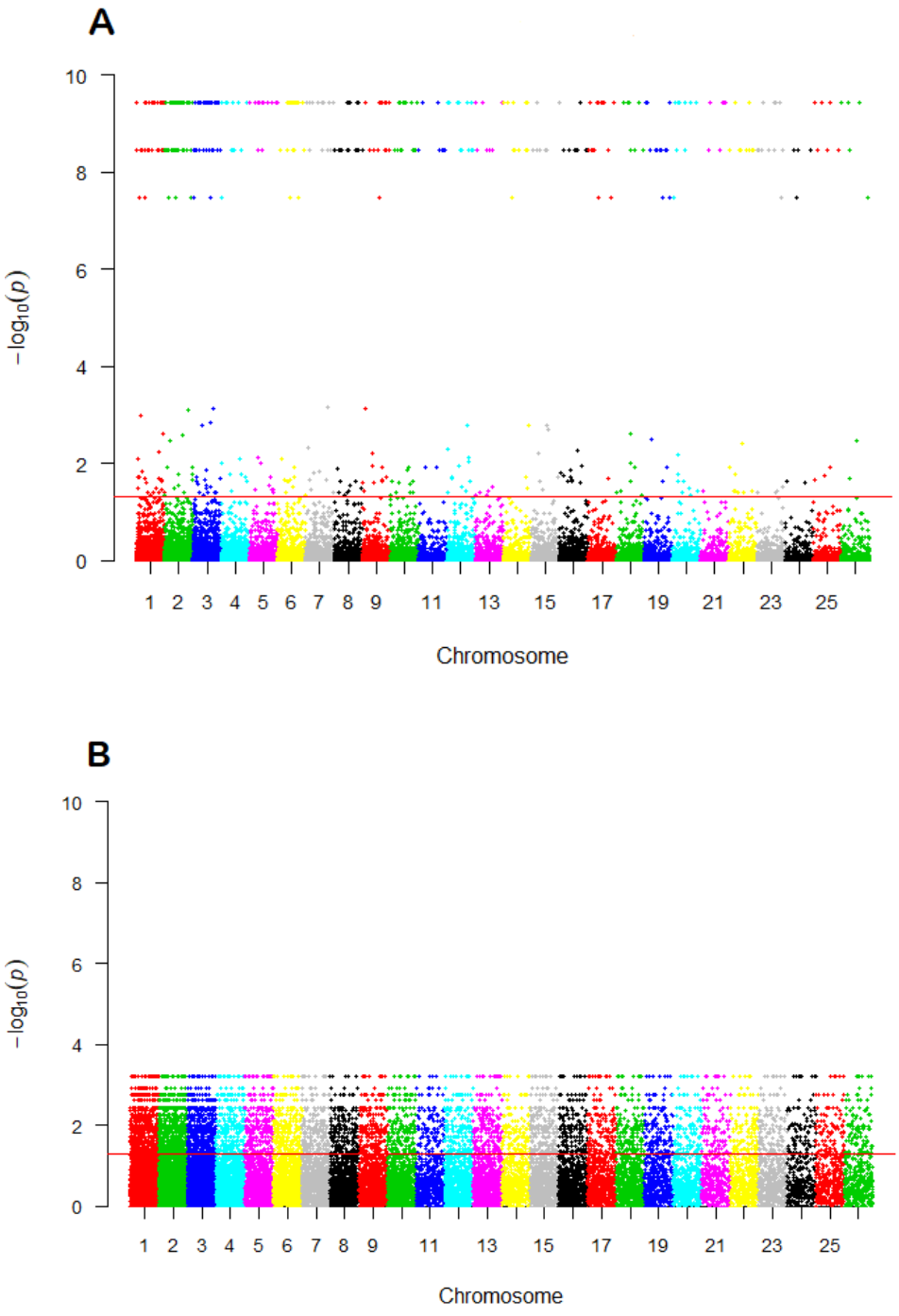

Figure **5-4**.: Manhattan plot of the SNPs associated with the hardness of meet in sheep detected using Benjamini and Hochberg´s proposal **(A)** and by Quotient C **(B)**. The red solid line represents $p = 0.05$.

SNPs	Position in the genome	X_i	r^*	$p-$values					Quotient C	
				Raw	Bonferroni	Holm	Hochberg	B-H	C	$p-$value
SNP_{63}	36268411	7.46	1	0.01	1	1	1	0.11	7459833	0.016
SNP_{79}	130749136	5.45	1	0.03	1	1	1	0.25	5448325	0.045
SNP_{93}	165282616	5.56	1	0.02	1	1	1	0.24	5563313	0.039
SNP_{126}	266680786	5.85	1	0.025	1	1	1	0.22	5856995	0.034
SNP_{180}	104167634	7.77	1	0.009	1	1	1	0.095	7771076	0.012
SNP_{22810}	47858558	37.44	9	*	*	*	*	*	264.67	*
SNP_{15399}	55183938	41.88	12	*	*	*	*	*	294.82	*
SNP_{13913}	49281284	46.56	14	*	*	*	*	*	355	*
SNP_{12679}	14041065	51.49	17	*	*	*	*	*	400.91	*
SNP_{11523}	78896526	56.67	21	*	*	*	*	*	464.02	*
SNP_{10253}	113158996	62.09	24	*	*	*	*	*	510.59	*
SNP_{9042}	189678502	67.77	28	*	*	*	*	*	589.82	*
SNP_{7876}	152010768	73.69	31	*	*	*	*	*	666.95	*
SNP_{30775}	158820235	79.86	36	*	*	*	*	*	739.43	*
SNP_{5282}	93548584	86.28	40	*	*	*	*	*	816.7	*
SNP_{3756}	204118069	92.95	45	*	*	*	*	*	914.71	*
SNP_{2366}	155320166	99.86	50	*	*	*	*	*	1018.5	*
SNP_{1122}	235621637	107.02	55	*	*	*	*	*	1117.55	*
SNP_{34}	244475780	114.44	60	*	*	*	*	*	1228.7	*
Abbreviations: *, p<0.0001; B-H, Benjamini and Hochberg´s method.										

Table **5-3**.: Some of the SNPs associated with the hardness of meet in sheep [48]. The SNPs were sorted in ascending order of their r^* value.

5.5.2. Statistical power

Table (**5-4**) reports the statistical power according to K as well as adjustment method. It is possible to see that when K is less than 2606, statistical power of all the methods is similar and close to 1. However, Bonferroni, Holm and Hochberg´s procedures showed an abrupt descent of power for $K \geq 2607$ (the number of SNPs detected as associated by these methods). Benjamini & Hochberg´s correction had a higher statistical power compared to the other classical methods, even for $K \geq 2795$ and its value decreased to 0 when $K = 2959$ (number of SNPs associated with the outcome using this procedure), showing a less abrupt decline than Bonferroni´s´s, Holm´s and Hochberg´s procedures.

When Quotient C was applied to detect SNPs associated with the outcome, the statistical power was higher than the other methods for multiple comparisons. Compared to Bonferroni´s, Holm´s and Hochberg´s proposals, when the latter ones reached zero power, Quotient C had values close to 1, even for 3787 SNPs ($\hat{\zeta} = 0.604$). The values of $\hat{\zeta}$ for Benjamini & Hochberg´s correction were lower in all the Ks compared to Quotient C; thus, when $\hat{\zeta} = 0$ in the Benjamini and Hochberg´s one, the statistical power for Quotient C was close to 0.817. For Quotient C, $\hat{\zeta} = 0$ when $K = 15267$, and for $K = 3787$, the probability of making a type II error was 0.396 (which implies that every 100 false null hypothesis, about 40 are erroneously not rejected).

Figure (**5-5**) shows the values of $\hat{\zeta}$ for each of the methods for different values of K.

5.6. Application of the proposed methodology to the dataset of SNPs from HLA-DRB1 associated with multiple sclerosis in an admixed Colombian population.

5.6.1. General results

Table (**5-5**) reports the raw and adjusted $p-$values as well as the value of r^*, C and p for Quotient C for each one of the SNPs studied. The Graph (**5-6**) shows a deviation of the raw $p-$values of the dotted line, which means a non-uniform profile of them and is possible to suspect that many of the tested hypothesis are false.

In an unadjusted analysis, two SNPs (DRB*15 and DRB1*14) were associated with multiple sclerosis (MS); when a classical correction was applied, only DRB1*15 was associated with the outcome in all the multiple testing procedures. Quotient C identified only the same SNPs as associated with MS ($p = 0.047$).

K	Statistical power				
	Bonferroni	Holm	Hochberg	B-H	Quotient C
1 - 600	1.0	1.0	1.0	1.0	1.0
601 - 1000	0.999	0.999	0.999	1.0	1.0
1001 - 2586	0.999	0.999	0.999	0.999	1.0
2587 - 2606	0.998	0.999	0.999	0.999	1.0
2607	0.0	0.0	0.0	0.991	1.0
2608	0.0	0.0	0.0	0.990	0.984
2666	0.0	0.0	0.0	0.805	0.984
2722	0.0	0.0	0.0	0.708	0.922
2746	0.0	0.0	0.0	0.607	0.922
2787	0.0	0.0	0.0	0.509	0.893
2815	0.0	0.0	0.0	0.411	0.878
2844	0.0	0.0	0.0	0.320	0.851
2902	0.0	0.0	0.0	0.216	0.838
2930	0.0	0.0	0.0	0.100	0.825
2959	0.0	0.0	0.0	0.0	0.817
3420	0.0	0.0	0.0	0.0	0.700
3787	0.0	0.0	0.0	0.0	0.604
4275	0.0	0.0	0.0	0.0	0.500
5160	0.0	0.0	0.0	0.0	0.403
5482	0.0	0.0	0.0	0.0	0.3216
7239	0.0	0.0	0.0	0.0	0.209
10848	0.0	0.0	0.0	0.0	0.101
15267	0.0	0.0	0.0	0.0	0.0
Abbreviations: B-H, Benjamini and Hochberg´s method.					

Table **5-4**.: Statistical power according to values of K and the different methods for the dataset of Ariza et al. [48].

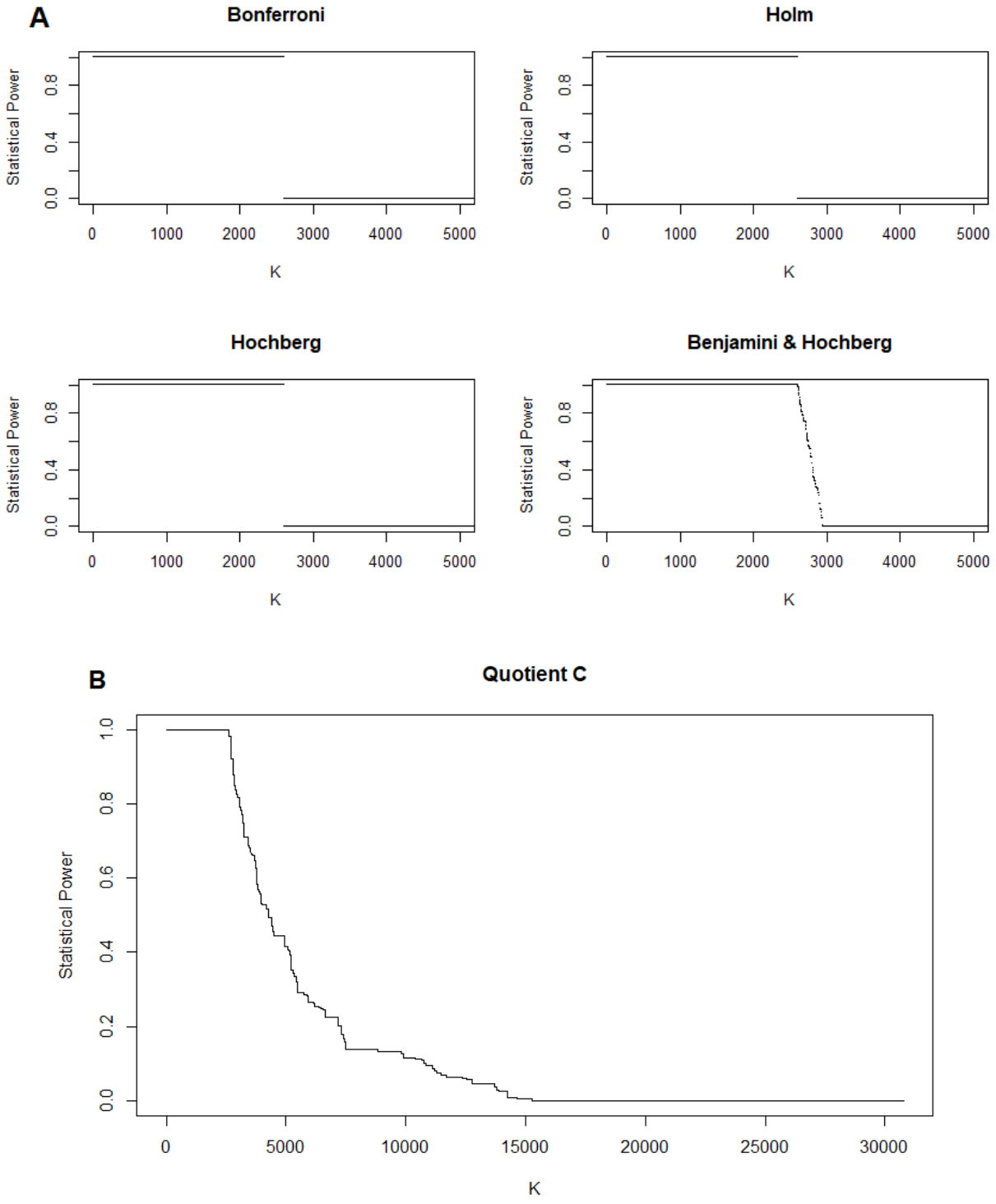

Figure **5-5**.: Line graphs of the statistical power for classical methods **(A)** and for Quotient C **(B)** in the dataset of Ariza et al. [48]

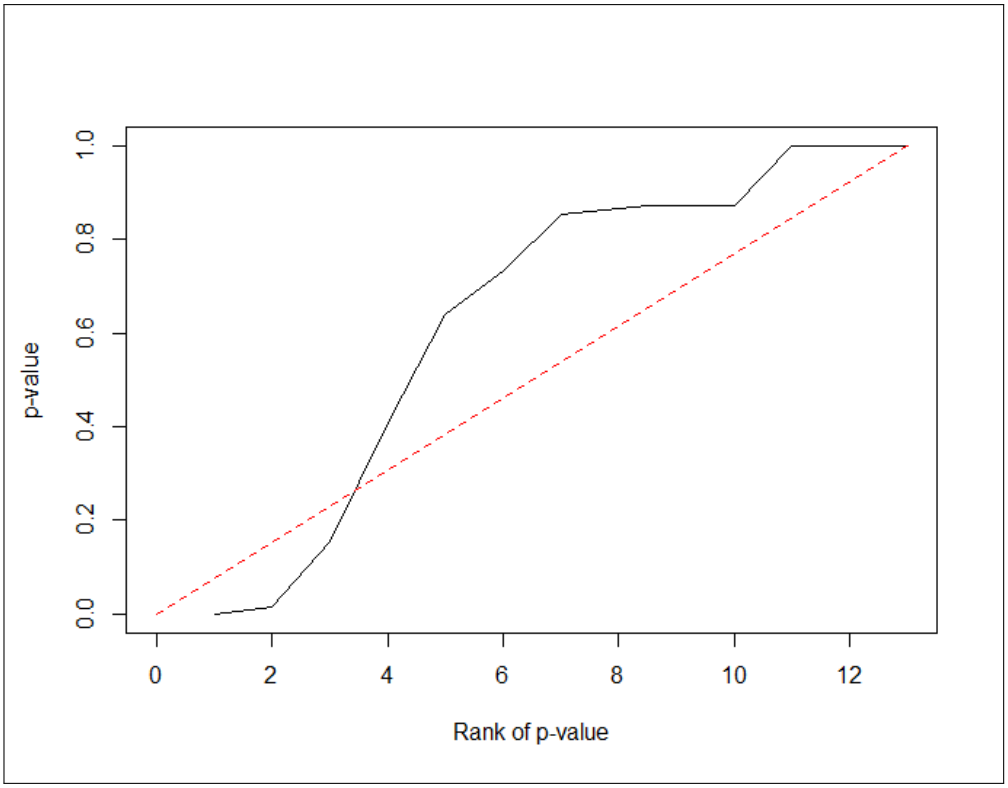

Figure **5-6**.: Profile of the raw $p-$values in the dataset of Toro et al. [50]

SNP	X_i	r^*	$p-$values					Quotient C	
			Raw	Bonferroni	Holm	Hochberg	B-H	C	$p-$value
DRB1*16	0.02105	1.00	1.00	1.00	1.00	1.00	1.00	107040452.17	0.78
DRB1*15	13.09133	7.00	0.00053	0.00694	0.00694	0.00694	0.00694	101.45	0.047
DRB1*14	6.96022	3.00	0.01429	0.18582	0.17153	0.17153	0.09291	41.10	0.34
DRB1*13	0.24021	1.00	0.73153	1.00	1.00	1.00	1.00	3211.13	0.67
DRB1*12	0.40677	1.00	0.87331	1.00	1.00	1.00	1.00	448.69	0.76
DRB1*11	2.60507	1.00	0.15335	1.00	1.00	1.00	0.66452	11.52	0.48
DRB1*10	0.06421	1.00	1.00	1.00	1.00	1.00	1.00	788973.34	0.82
DRB1*09	0.40677	1.00	0.87331	1.00	1.00	1.00	1.00	448.69	0.85
DRB1*08	1.03504	1.00	0.40845	1.00	1.00	1.00	1.00	28.74	0.80
DRB1*07	0.10162	1.00	0.87335	1.00	1.00	1.00	1.00	109669.72	0.68
DRB1*04	0.00682	1.00	1.00	1.00	1.00	1.00	1.00	16413046735.95	0.69
DRB1*03	0.38986	1.00	0.63957	1.00	1.00	1.00	1.00	521.62	0.87
DRB1*01	0.13262	1.00	0.85551	1.00	1.00	1.00	1.00	35728.89	0.82

Abbreviations: B-H, Benjamini and Hochberg´s method.

Table **5-5**.: Values of X_i, r^* and p according to adjustment methods and values of Quotient C for the SNPs in the dataset of Toro et al. [50]

K	Method				
	Bonferroni	Holm	Hochberg	B-H	Quotient C
1	0.56	0.56	0.56	0.56	0.21
2	0.00	0.00	0.00	0.04	0.21
≥ 3	0.00	0.00	0.00	0.00	0.00
Abbreviations: B-H, Benjamini and Hochberg´s method.					

Table **5-6**.: Statistical power according to values of K and multiple testing methods in the dataset of Toro et al. [50]

5.6.2. Statistical power

The statistical power of the classical multiple testing procedures for $K = 1$ was similar for all of them and higher than the Quotient C one. So, for $K = 1$, the probability of rejecting the null hypothesis when there are association between the SNP and the outcome is equal to 0.56 when Bonferroni´s, Holm´s, Hochberg´s and Benjamini and Hochberg´s corrections were applied and 0.21 for Quotient C. When $K = 2$, $\hat{\zeta} = 0$ for classical methods but Quotient C statistical power remained equal when $K = 1$. The value of $\hat{\zeta}$ was 0 for Quotient C when $K \geq 3$ (see Table (**5-6**)).

5.7. Simulations

The table (**5-7**) reports the results of the performed simulation using the dataset of Single Nucleotide Polymorphism associated with hardness of meat in sheep. Among the sample of SNPs detected by all the methods (782), all the compared procedures were able to detect the 100% of them in each one of the 1000 simulations. On the other hand, in the sample of polymorphisms detected only by Benjamini and Hochberg, the Bonferroni´s procedure did not find as associated any of them and the Quotient C detected as associated approximately 17-18 of those SNPs in each one of the simulations. Finally, the Quotient C identified during the simulations nearly 305-308 of the 308 SNPs detected only by this proposal, meanwhile the other two methods did not find any as associated with the outcome. (5-11)

The simulation study has shown that Quotient C and Benjamini and Hochberg´s proposal are consistent with respect to the SNPs that are identified as associated by themselves. The number of polymorphism associated was always equal to the number of the sample for both methods. In addition, Quotient C identified in 1000 repetitions, SNPs that Benjamini and Hochberg´s method did not in contrast with the first methodology that identified some polymorphism associated by Benjamini and Hochberg.

	SNPs detected by all the methods (Size: 782)			SNPs detected only by B-H (Size: 57)			SNPs detected only by Quotient C (Size: 308)		
	SNPs detected by								
	Bonferroni	B-H	Quotient C	Bonferroni	B-H	Quotient C	Bonferroni	B-H	Quotient C
Mean (SD)	782 (0)	782(0)	782 (0)	0(0)	57(0)	17(3)	0(0)	0(0)	305(11)
Median (IR)	782 (782-782)	782(782-782)	782(782-782)	0(0-0)	57(57-57)	18(16-20)	0(0-0)	0(0-0)	308 (308-308)

Abbreviations: B-H, Benjamini and Hochberg´s method; SD, standard deviation; IR, interquartile range.
The results are based on 1000 simulations.

Table **5-7**.: Simulation results

6. Discussion

The classical methods (Bonferroni, Holm and Hochberg) control the multiple testing by the control of the *familywise error rate* (FWER) given by $FWER = P(V > 0) = P(Q > 0)$ (probability that $\mathcal{R}$ contains any error) [41, 54, 55, 98]. Meanwhile Benjamini and Hochberg´s procedure controls the *false discovery rate* (FDR) or $E(Q)$ (the expected proportion of errors among $\mathcal{R}$) [56]. For those methods, the researcher selects an error rate to control type I error and the procedure finds a collection $\mathcal{R}$ according to an *a priori* criterion, (generally the significance level, or α). This procedure is straightforward regarding the control of false positives but affects power enormously [90]. Moreover, given the amount of statistical tests done in genomic data, the procedure of multiple testing correction is not the best choice.

As shown, the here proposed test statistic, Quotient C, solves several problems of previous procedures as it allows to correct type I error and increases power at the same time, being simple to apply to SNP data and a categorical trait. In the dataset of Ariza et al. [48], the Bonferroni´s, Holm´s and Hochberg´s procedure detected a lower amount of SNPs associated with the outcome than Benjamini and Hochberg´s method. This amount was lower than the number of polymorphism detected by Quotient C. The FWER-based methods are considered very strict and therefore very conservative at level α because the probability of wrongly rejecting a true null hypothesis is $\leq \alpha$ [60, 91, 92, 62]. Controlling the *familywise error rate* implies to prevent type I errors at all cost (thus, when the FWER-based procedures are used, it results in very few rejected hypotheses or non at all, but the individually rejected hypotheses are reliable) and consequently the amount of type II errors are high [41, 62, 93, 94]. Those type of errors can be very expensive in a clinical, biological or animal breeding context. If a type II error is done and false negatives appear, this means that a disease is not detected, or in the genomic context, that a SNP is declared as non associated with a particular trail when it is really associated with it. Although Bonferroni´s, Holm´s and Hochberg´s method control the same type of error, the last two are considered less conservative [93, 95], perhaps in the two datasets used, the number of rejected null hypotheses was equal in all the FWER-based methods. The Benjamini and Hochberg´s proposal is considered less conservative than FWER-controlling methods [96, 97] and its critical values are much larger, so usually more rejections can be made; in the dataset of the hardness of the meat, 189 more SNPs were detected as associated with a outcome than FWER´s methods.

In FWER approach, if the set of hypotheses $\mathcal{R}$ has $FWER \leq \alpha$, each one of those hypotheses in $\mathcal{R}$ has a probability $\leq \alpha$ of being a type I error, meanwhile, in FDR control methods, the collection $\mathcal{R}$ has *on average* a probability α on that subset (not each individual hypothesis) [41, 98, 99]. That is an advantage of FWER-based procedures over FDR-based procedures. Although the last one is considered one of the best options in multiple testing problems, its usefulness is still being criticized

due to the property of the controls and the implications of its conclusions. An open question is how risky it can be to draw conclusions using the individually rejected hypotheses [41, 100, 111]. As FDR= $E[Q]$, the collection $\mathcal{R}$ can have FDP> α. That implies that controlling the FDR at an given level of α can allow that the amount of false rejecting on $\mathcal{R}$ is greater than the type I error allowed in the study. Since the set $\mathcal{R}$ has on average FDP$\leq \alpha$, this measure is not the same for any sub collection on $\mathcal{R}$, which can have higher or less values of FDP than the complete collection. Thus, is likely that $\mathcal{R}$ has false positives and an individually rejected hypothesis can be a false positive indeed [41, 102, 108]. A subsequent disadvantage of this situation is the interpretation of the adjusted p values; since they reflect a property of the entire collection $\mathcal{R}$ and not of each individual hypothesis. Therefore decisions on individual SNPs are risky.

For both data sets, all the SNPs associated using Benjamini and Hochberg´s method were the same detected by Bonferroni´s, Holm´s and Hochberg´s one. The FWER and FDR are related: how $0 \leq \frac{V}{R} \leq 1$, then $FDR \leq P(Q > 0)$, which implies that all the methods that control the FWER are also methods that control FDR [41].

Since $E(Q) \leq P(Q > 0)$, FDR is at the most equal to the FWER and it is easier to keep the FDR below α than to keep the FWER at the same level. So, the expect result is that FDR-controlling procedures have more power compared to FWER-based methods. To keep the level α has an important impact over the errors committed; FWER-controlling methods introduce a greater amount of type II errors, so its statistical power is lower compared to methods that control the *false discovery rate*. Benjamini and Hochberg´s proposal is considered the most powerful procedure in multiple testing problems because the FDR control is less stringent than FWER-controlling methods [104, 100, 105]. In both data sets, the statistical power was similar for the classical methods until a number of SNPs equal to the number of SNPs detected as associated with the outcome was reached for FWER-controlling methods, when $K > \#\mathcal{R}$, the Benjamini and Hochberg´s statistical power was higher even beyond the number of SNPs declared as associated by this proposal. Relaxing the criterion from FWER to FDR and then to Quotient C, clearly makes an important difference in terms of improving statistical power.

We propose two applications of the defined collection $\mathcal{R}$ using the Quotient C: (i) to be a test statistic by itself to evaluate the association between SNPs and an outcome at a level α, allowing to generate conclusions from individually rejected hypothesis (this a consequence of being a FWER-controlling method) and (ii) in a conservative way, to define an exploratory set of hypothesis to be tested later using classical multiple comparisons procedures (be used as a preliminary filter). In both cases, a less stringent condition in which $H_k, k = 1, 2, \ldots, K$ are associated if it belongs to the maximum of the distribution has shown to improve the statistical power (and to detect a higher amount of SNPs associated) compared to the classical methods, principally compared to those that control FWER.

In the Quotient C proposal, no assumptions beyond independence between the hypothesis tests were assumed (not over the $p-$values). In multiple testing problems, mainly in genomics, some dependence structure of the $p-$values must be considered (no dependence, *positive dependence through stochastic ordering* and some methods use an adapted dependency structure of the $p-$values using permutations) [41]. Generally, the data show a degree of internal correlation as statistical

dependencies between variables [89, 106]. Bonferroni´s and Holm´s proposal are valid under any dependence structure of the $p-$values, Hochberg is valid if the PDS assumption holds as well as Benjamini and Hochberg [41, 107, 108]. The worst scenario about the dependency structure was considered in our proposal and more conservative results were obtained with respect to other methods that consider different dependence structure of the $p-$values. However, future studies must be performed to establish how the proposed statistical test works regarding other assumptions over the $p-$values. Ignoring the dependence between the hypothesis test (and their respective $p-$values) can result in both bias due to the priori unknown correlation or highly variable significance measures [57, 109]. Other authors [106] have shown that the correlation may increase the bias and variance of the false discovery rate estimation substantially with respect to the independent case, and that in some cases, such as an exchangeable correlation structure, the estimator fails to be consistent as the number of tests becomes large. The current consensus establishes that, in genomic data, it is safe to use methods based on some structure dependence of the $p-$values [41, 110, 111]. From the profile of the raw $p-$values in both datasets, some doubts exist about the dependence structure of they; although a uniformity of the p-values of true hypotheses is expected on average, the existence of correlation between them can explain the shape of those figures [41, 112]. This is true, especially for the data set of HLA-DRB1 associated with multiple sclerosis in an admixed Colombian population. In GWAS, usually very few SNPs are expected as associated with the outcome and therefore a lower amount of false null hypothesis, therefore, a deviation from the uniform shape of the histogram or from the dotted line on the sorted $p-$values is taken also as an indication of lack of model fit or of inappropriate formulation of null hypotheses [41, 113, 114]. Without further analysis, it is not possible to attribute the deviations seen from uniformity to the presence of some dependence structure of the $p-$values or false hypotheses.

The statistical problem that was addressed implies to find a new method to establish associations between exposures (alleles of SNPs) and one dichotomous outcome. This univariate view of the problem is a situation that is not completely realistic. In most of the diseases, as in complex diseases, only a genetic variation is not enough to explain an outcome. Behavioral and environmental factors mediate the way in which a genetic variant is expressed to produce a particular trait. In other situations, the outcome is a result of interactions between genetic variations and external factors. A broader approach to the statistical and biological problem requires a multivariate view; Quotient C allows a univariate solution to this particular problem in *Genome-wide association studies* and further work should be addressed to reach a multivariate extension of the proposed method, controlling the effect of multiple independent variables (including interaction between SNPs) on the outcome.

From the expressions (5-7) it is possible to see that when K, the number of tested SNPs is too large, a common situation in GWAS, the calculus of $f_C(c)$ is problematic. *Genome-wide association studies* often involve hundreds, thousands or even millions of SNPs tested simultaneously, so K is expected to be large and computationally, the implementation of Quotient C can be limitated. This limitation has a clear impact on the utility of Quotient C in terms of precision and acceptance as a alternative for the scientific community and a more general public.

Although Quotient C was originally designed to solve a particular statistical problem deviated from a biologic problem, its usefulness is able to be extended to other scenarios in which associations

between dichotomous variables must be found in the context of a multiple testing problem as social sciences, Economy, Medicine, Epidemiology, Agronomy and animal health, among other. In my work, I have shown that the classical methods are not the only solution to this type of situations. Some of the most important questions about Quotient C are now open and future work must be performed to understand clearly the limits of my proposal.

7. Conclusions and perspectives

- Quotient C represents a new option to solve the problems generated in multiple testing problems, with a direct application to GWAS and other scenarios. It represents a less stringent criterion and allows to declare a larger amount of associations between SNPs and dichotomous outcomes compared to the classical methods used to correct for multiple testing, keeping the probability of incorrectly rejecting a true null hypothesis (type I error) equal or less than α.

- The proposed methodology has a lower type II error rate and a better statistical power than Bonferroni´s, Holm´s, Hochberg´s and Benjamini & Hochberg´s methods. To improve the statistical power without increasing the sample size and without forgetting the high-dimensionality of the problem represents clear advantages of Quotient C compared to the previous proposals, allowing minimizing the amount of false negatives on the results. Also, the procedure is simple to implement and its interpretation is easy for statisticians as well as for scientists that produce the data.

- Future works must be performed in order to establish the validity of the proposed methodology under other assumptions and to expand the proposal so that a broader approach to the statistical and biological problem will be reached.

8. Ethical considerations

This thesis was developed following the *Ethical Guidelines for Statistical Practice* Prepared by the *Committee on Professional Ethics of the American Statistical Association*[115]. (8-0)

For the development and applying of the proposed methodology, there were not participant recruitment or obtaining biological samples of any type. In addition, there were no sacrifice of animals. So, to reach the proposed objectives was used data sets given by Ariza et al. [48] and Toro et al. [50]. (8-1)

The research *Evaluación genómica asociada a características de calidad de la carne de ovino de pelo criollo colombiano* was approved by Comité de Bióetica of the Faculty of veterinary medicine and zootechnics at the Universidad Nacional de Colombia. Likewise, the research *HLA-DRB1*14 is a protective allele for multiple sclerosis in an admixed Colombian population* was approved by Comité Corporativo de Ética en Investigación at Fundación Santa Fé de Bogotá. Both committees warranted that the original research that recollected the data in each one of the data sets used in this thesis complied with ethical principles to conduct research in animals and humans according to each case.

Bibliography

[1] Duncan, DB. Bayes rules for a common multiple comparisons problem and related. Student-t problems. Ann Math Statist 1961; 32: 1013-1033.

[2] Duncan, DB. A Bayesian approach to multiple comparisons. Technometrics 1965; 7: 171-222.

[3] Tukey, J. W. The Problem of Multiple Comparisons. Mimeographed monograph (1953).

[4] Benjamini, Y and Braun, H. John W. Tukey's contributions to multiple comparisons. The Annals of Statistics 2002; Vol. 30, No. 6: 1576-1594.

[5] The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001; 409: 928-33.

[6] Shastry, B.S. SNP alleles in human disease and evolution. J. Hum. Genet. 2002; 47: 561-566.

[7] Brookes, A.J. The essence of SNPs. Gene 1999; 234: 177-186.

[8] William Y. S. Wang, Bryan J. Barratt, David G. Clayton, and John A. Todd, Genome-Wide Association Studies: Theoretical and Practical Concerns. Nat Rev Genet. 2005 Feb;6(2): 109-18.

[9] Dudley JT, Kim Y, Liu L, et al. Human genomic disease variants: A neutral evolutionary explanation. Genome Research. 2012; 22(8): 1383-1394.

[10] Manolio TA. Implementing genomics and pharmacogenomics in the clinic: The National Human Genome Research Institute´s genomic medicine portfolio. Atherosclerosis. 2016;253: 225-236.

[11] Hurle B, Citrin T, Jenkins JF, et al. What Does it Mean to be Genomically Literate? National Human Genome Research Institute Meeting Report. Genetics in medicine: official journal of the American College of Medical Genetics. 2013; 15(8): 658-663.

[12] Moyses Szklo and F. Javier Nieto. Epidemiology: Beyond the Basics. 3rd Ed, Jones / Bartlett Learning, 2014.

[13] Gordis, Leon. Epidemiology. 5th Ed, Elsevier, 2014.

[14] Fletcher W, Roberth & Fletcher W, Suzanne. Clinical Epidemiology, the essentials. 4th Ed, Lippincott Williams Wilkins, 2005.

[15] Kenneth J. Rothman, Timothy L. Lash & Sander Greenland. Modern Epidemiology. 3rd Ed, Lippincott Williams Wilkins, 2012.

[16] Carlson, C. S., Eberle, M. A., Kruglyak, L. & Nickerson, D. A. Mapping complex disease loci in whole-genome association studies. Nature 2004; 429: 446-452.

[17] Weiss, K. M. & Terwilliger, J. D. How many diseases does it take to map a gene with SNPs? Nature Genet. 2000; 26: 151-157.

[18] Drakoulis N, Cascorbi I, Brockmoller J, Gross CR, Roots I. Polymorphisms in the human CYP1A1 gene as susceptibility factors for lung cancer: exon-7 mutation (4889 A to G), and a T to C mutation in the 3'-flanking region. Clin Investig. 1994 Feb; 72(3): 240-8.

[19] Rita M. Cantor, Kenneth Lange and Janet S. Sinsheimer. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics 2010; 86: 6-22.

[20] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (April 2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science. 2005 Apr 15; 308(5720): 385â“389.

[21] Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature Genet. 2003; 33: 177-182.

[22] Benjamin Brachi, Geoffrey P Morris and Justin O Borevitz. Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol. 2011 Oct 28; 12(10): 232.

[23] Claudia Bartoli and Fabrice Roux. Genome-Wide Association Studies In Plant Pathosystems: Toward an Ecological Genomics Approach. Front Plant Sci. 2017; 8: 763.

[24] Bernadette Rubio, Patrick Cosson, Mélodie Caballero, Frédéric Revers. Joy Bergelson, Fabrice Roux and Valérie Schurdi-Levraud. Genome wide association study reveals new loci involved in Arabidopsis thaliana and Turnip mosaic virus (TuMV) interactions in the field. New Phytol. 2019 Mar; 221(4): 2026-2038.

[25] Hui Zhang, Zhipeng Wang, Shouzhi Wang and Hui Li. Progress of genome wide association study in domestic animals. J Anim Sci Biotechnol. 2012 Aug 22;3(1): 26.

[26] Aditi Sharma, Jun Seop Lee, Chang Gwon Dang, Pita Sudrajad, Hyeong Cheol Kim, Seong Heum Yeon, Hee Seol Kang, and Seung-Hwan Lee. Asian-Australas J Anim Sci. Stories and Challenges of Genome Wide Association Studies in Livestock - A Review. Asian-Australas J Anim Sci. 2015 Oct; 28(10): 1371-9.

[27] Ian W. Richardson, Donagh P. Berry, Heather L. Wiencko, Isabella M. Higgins, Simon J. More, Jennifer McClure,David J. Lynn, and Daniel G. Bradley. A genome-wide association study for genetic susceptibility to Mycobacterium bovis infection in dairy cattle identifies a susceptibility QTL on chromosome 23. Genet Sel Evol. 2016 Mar 9; 48: 19.

[28] Mohammed K. Abo-Ismail, Luiz F. Brito, Stephen P. Miller, Mehdi Sargolzaei, Daniela A. Grossi, Steve S. Moore, Graham Plastow, Paul Stothard, Shadi Nayeri and Flavio S. Schenkel. Genome-wide association studies and genomic prediction of breeding values

for calving performance and body conformation traits in Holstein cattle. Genet Sel Evol. 2017 Nov 7; 49(1): 82.

[29] Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. Nat Genet. 2001 Mar; 27(3): 234-6.

[30] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1; 29(1): 308-311.

[31] SeattleSNPs PGA. Program of Genomica Appllications supported by the National Heart, Lung and Blood Institute.

[32] EVOLution of TREEs as drivers of terrestrial biodiversity. Retrieved October 24, 2018 from: http://www.evoltree.eu/index.php/e-recources/databases

[33] James E. Gentle. A companion for mathematical statistics. Fairfax County, Virginia, 2010.

[34] Robert V Hogg, Joseph W McKean, Allen T Craig. Introduction to Mathematical Statistics. Sixth Edition. Pearson Education International, 2005.

[35] Shao, Ju. Mathematical Statistics. Springer texts in Statistics. 2nd ed. 2003.

[36] Johnson et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics. 2010; 11: 724.

[37] Storey JD and Tibshirani R. Statistical significance for genome-wide studies. Proceedings of the National Academy of Science 2003, 100(16): 9440-9445.

[38] Spencer CCA, Su Z, Donnelly P, Marchini J. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. PLoS Genet 2009; 5(5): e1000477. https://doi.org/10.1371/journal.pgen.1000477

[39] Han B, Kang HM, Eskin E. Rapid and Accurate Multiple Testing Correction and Power Estimation for Millions of Correlated Markers. PLoS Genet 2009; 5(4): e1000456. https://doi.org/10.1371/journal.pgen.1000456

[40] Lon R. Cardon and John I. Bell. Association Study Designs For Complex Diseases. Nat Rev Genet. 2001 Feb;2(2): 91-9.

[41] Goeman JJ & Solari A. Multiple hypothesis testing in genomics. Stat Med. 2014 May 20; 33(11): 1946-78.

[42] Brent, Richard P. Algorithms for minimization without derivatives. Englewood Cliffs, N.J. Prentice-Hall, 1973.

[43] One Dimensional Optimization [Internet]. R-manual [cited 2018 Nov 26]. Available from: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optimize.html

[44] Sandrine Dudoit, Juliet Popper Shaffer & Jennifer C. Boldrick. Multiple Hypothesis Testing in Microarray Experiments. Statistical Science 2003; Vol. 18, No. 1: 71-103.

[45] Abramowitz, Milton. Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables. Dover Publications, Inc. New York, NY, USA.

[46] Lehmann, E.L. and Romano, J. P. (2005). Testing Statistical Hypotheses. (Third Ed), London: Springer.

[47] Dudoir, S., Shaffer, J. P. and Boldrick, J. C. (2002). Multiple hypothesis testing in microarray experiments. Technical Report 110, Division of Biostatistics, Univ. California, Berkeley. Available at http://www.bepress.com/ucbbiostat/paper110/.

[48] Ariza, Manuel Fernando & Grupo de Investigación en Genética Molecular Animal de la Universidad Nacional de Colombia. Evaluación genómica asociada a características de calidad de la carne de OPCC. Proyecto financiado por CENIRED y Ministerio de Agricultura y Desarrollo Rural de Colombia, 2018.

[49] Ilumina, Inc. OvineSNP50 Genotyping BeadChip. Information avaible in: https://www.illumina.com/documents/products/datasheets/datasheet_ovinesnp50.pdf.

[50] Toro J, Cuellar-Giraldo D, DÃaz-Cruz C, et al. HLA-DRB1*14 is a protective allele for multiple sclerosis in an admixed Colombian population. Neurol Neuroimmunol Neuroinflamm. 2015 Dec 24;3(1):e192. doi: 10.1212/NXI.0000000000000192. eCollection 2016 Feb.

[51] Mc Intyre LM, Martin ER, Simonsen KL, Kaplan NL. Circumventing multiple testing: A multilocus Monte Carlo approach to testing for association. Genet Epidemiol 2000; 19:18-29.

[52] Stephen Olejnik, Jianmin Li, Suchada Supattahum and Carl J. Huberty. Multiple Testing and Statistical Power With Modified Bonferroni Procedures. Journal Of Educational And Behavioral Statistics. December 21, 1997; vol. 22 No.4: 389-406.

[53] Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 1967; 62(318): 626-633.

[54] Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 1979; 6(2): 65-70.

[55] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988; 75(4): 800-802.

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

[57] Jeffrey T. Leek and John D. Storey. A General Framework for Multiple Testing Dependence. Proceedings of the National Academy of Sciences of the United States of America. Vol. 105, No. 48 (Dec. 2, 2008), pp. 18718-18723.

[58] Helmut Finner, Thorsten Dickhaus, & Markus Roters. Dependency and false discovery rate: Asymptotics. Annals of Statistics 2007; Vol. 35, No. 4: 1432-1455.

[59] Sanat K.Sarkar. Two-stage stepup procedures controlling FDR. Journal of Statistical Planning and Inference 2008; Volume 138, Issue 4: 1072-1084.

[60] John D. Storey. A direct approach to false discovery rates. J. R. Statist. Soc. B 2002; 64, Part 3: 479-498.

[61] Efron B, Tibshirani R, Storey J, Tusher V. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 2001; 96(456): 1151-1160.

[62] Jelle J. Goeman and Aldo Solari. Multiple Testing for Exploratory Research. Statist. Sci. 2011;Volume 26, Number 4: 584-597.

[63] Meinshausen N. False discovery control for multiple tests of association under general dependence. Scandinavian Journal of Statistics 2006; 33(2): 227-237.

[64] Austin E, Pan W, Shen X. Penalized Regression and Risk Prediction in Genome-Wide Association Studies. Stat Anal Data Min. 2013;6(4): 10.1002/sam.11183.

[65] Tibshirani, R. J. Regression Shrinkage and Selection via the LASSO. Journal of the Royal Statistical Society 1996; Ser. B, 58: 267-288.

[66] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of American Statistical Association. 2001; 96: 1348-1360.

[67] Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. Journal of American Statistical Association. 2012; 107: 223-232.

[68] Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price AL. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet. 2019 Jan 3;104(1): 65-75.

[69] Fernando R.L., Garrick D. (2013) Bayesian Methods Applied to GWAS. In: Gondro C., van der Werf J., Hayes B. (eds) Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols), vol 1019. Humana Press, Totowa, NJ.

[70] Verzilli CJ, Stallard N, Whittaker JC. Bayesian graphical models for genomewide association studies. Am J Hum Genet. 2006; 79(1): 100-12.

[71] Zhang X, Xue F, Liu H, et al. Integrative Bayesian variable selection with gene-based informative priors for genome-wide association studies. BMC Genet. 2014 Dec 10; 15:130.

[72] Bing Han, Meeyoung Park and Xue-wen Chen. A Markov blanket-based method for detecting causal SNPs in GWAS. Selected articles from the 2009 IEEE International Conference on Bioinformatics and Biomedicine. BMC Bioinformatics 2010; 11(Suppl 3): S5.

[73] He Q, Lin DY. A variable selection method for genome-wide association studies. Bioinformatics. 2011; 27(1): 1-8.

[74] Laurent Briollais, Adrian Dobra , Jinnan Liu , Hilmi Ozcelik and Helene Massam. A Bayesian Graphical Model for Genome-wide Association Studies (GWAS). Ann. Appl. Stat. 2016; volume 10 number 2: 786-811.

[75] Zhang Y and Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nat Genet. 2007; 39(9): 1167-73.

[76] Wilson MA, Iversen ES, Clyde MA, Schmidler SC, Schildkraut JM. Bayesian model search and multilevel inference for SNP association studies. Ann Appl Stat. 2010; 4(3): 1342-1364.

[77] John P A Ioannidis, Thomas A Trikalinos, Evangelia E Ntzani, Despina G Contopoulos-Ioannidis. Genetic associations in large versus small studies: an empirical assessment. Lancet 2003; 361: 567-71.

[78] Loannidis JP, Cappelleri JC, Lau J. Issues in comparisons between metaanalyses and large trials. JAMA 1998; 279: 1089-93.

[79] Giovanni Gambaro, Franca Anglani, Angela D'Angelo. Association studies of genetic polymorphisms and complex Disease. Lancet 2000; 355: 308-11.

[80] Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. Am. J. Hum. Genet. 2001; 69: 936-950.

[81] Joel N. Hirschhorn and Mark J. Daly. Genome-Wide Association Studies For Common Diseases And Complex Traits. Nature Reviews Genetics 2005; 6: 95-108.

[82] Bonferroni, C. E., Teoria statistica delle classi e calcolo delle probabilitÃ . Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936.

[83] Shao, Jun. Mathematical Statistics. Springer, 2nd Edition, 2003.

[84] Dennis Wackerly , William Mendenhall & Richard L. Scheaffer. Mathematical Statistics with Applications. 7th ed, Thomson Learning 2008.

[85] Peter J. Bickel Kjell A. Doksum. Mathematical Statistics, Basic Ideas and Selected Topics. 2nd ed, Vol. 1. Prentice Hall, 2001.

[86] Jacod Jean & Protter Philip. Probability Essentials. Springer, 2004.

[87] Shaffer, J. P. Multiple hypothesis testing: A review. Annual Review of Psychology 1995; 46: 561-584.

[88] Finner H, Roters M. On the false discovery rate and expected type I errors. Biometrical Journal 2001; 43(8): 985-1005.

[89] William Poole, David L Gibbs, Ilya Shmulevich Brady Bernard and Theo A Knijnenburg. Combining dependent P-values with an empirical adaptation of Brownâ^{TM}s method. Bioinformatics 2016; Volume 32, Issue 17: i430â"i436.

[90] López-Kleine, L. and González-Prieto, C. Challenges Analyzing RNA-Seq Gene Expression Data. Open Journal of Statistics 2016; 6: 628-636.

[91] Keselman, H., Cribbie, R., Holland, B. Controlling the rate of Type I error over a large set of statistical tests. British Journal of Mathematical and Statistical Psychology 2002; 55: 27-39.

[92] Kvam, V., Liu, P., Si, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. American Journal of Botany 2012; 99(2): 248-256.

[93] Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical methods in medical research. 2003; 12(5): 419-446.

[94] Perneger T. What´s wrong with Bonferroni adjustments. British Medical Journal 1998; 316(7139): 1236-1238.

[95] Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Statistical Methods in Medical Research 2008; 17(4): 347-388.

[96] Roquain E. Type I error rate control for testing many hypotheses: a survey with proofs. Journal de la Societé Francaise de Statistique 2011; 153(2): 3-38.

[97] Dmitrienko A, D´Agostino R. Traditional multiplicity adjustment methods in clinical trials. Statistics in Medicine 2013; 32(29): 5172-5218.

[98] Finner H, Roters M. On the false discovery rate and expected type I errors. Biometrical Journal 2001; 43(8): 985-1005.

[99] Dudoit S, Van der Laan M. Multiple Testing Procedures with Applications to Genomics. Springer Verlag: New York, 2008.

[100] Korn E, Troendle J, McShane L, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. Journal of Statistical Planning and Inference 2004; 124(2): 379-398.

[101] Troendle J. Comments on: control of the false discovery rate under dependence using the bootstrap and subsampling. Test 2008; 17(3): 456-457.

[102] Blanchard G, Roquain E. Two simple sufficient conditions for FDR control. Electronic Journal of Statistics 2008; 2: 963-992.

[103] Romano J, Shaikh A, Wolf M. Control of the false discovery rate under dependence using the bootstrap and subsampling. Test 2008; 17(3): 417-442.

[104] Posch M, Zehetmayer S, Bauer P. Hunting for significance with the false discovery rate. Journal of the American Statistical Association 2009; 104(486): 832-840.

[105] Troendle J. Stepwise normal theory multiple test procedures controlling the false discovery rate. Journal of Statistical Planning and Inference 2000; 84(1-2): 139-158.

[106] Schwartzman A, Lin X. The effect of correlation in false discovery rate estimation. Biometrika 2011; 98(1): 199-214.

[107] Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference 1999; 82(1-2): 171-196.

[108] Romano J, Shaikh A, Wolf M. Control of the false discovery rate under dependence using the bootstrap and subsampling. Test 2008; 17(3): 417-442.

[109] Qiu X, Klebanov L, Yakovlev A. Correlation between gene expression levels and limitations of the empirical Bayes methodology in microarray data analysis. Statistical Applications in Genetics and Molecular Biology 2005; 4(1): 34.

[110] Ge Y, Sealfon S, Speed T. Some step-down procedures controlling the false discovery rate under dependence. Statistica Sinica 2008; 18(3): 881-904.

[111] Troendle J. Comments on: control of the false discovery rate under dependence using the bootstrap and subsampling. Test 2008; 17(3): 456-457.

[112] Schweder, T. and Spjotvoll, E. Plots of P-values to evaluate many tests simultaneously. Biometrika 1982; 69: 493-502.

[113] H. M. James Hung, Robert T. O'Neill, Peter Bauer and Karl Kohne Source. The Behavior of the P-Value When the Alternative Hypothesis is True Author(s). Biometrics 1997; Vol. 53, No. 1: 11-22.

[114] Parker, R. A. and Rothenberg, R. B. Identifying important results from multiple statistical tests. Statistics in Medicine 1988; 7: 1031-1043.

[115] Committee on Professional Ethics of the American Statistical Association. 2018. Ethical Guidelines for Statistical Practice. [online]. Available at: https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

A. Appendix: Plot function of $f_C(c)$ for different values of r^* and K

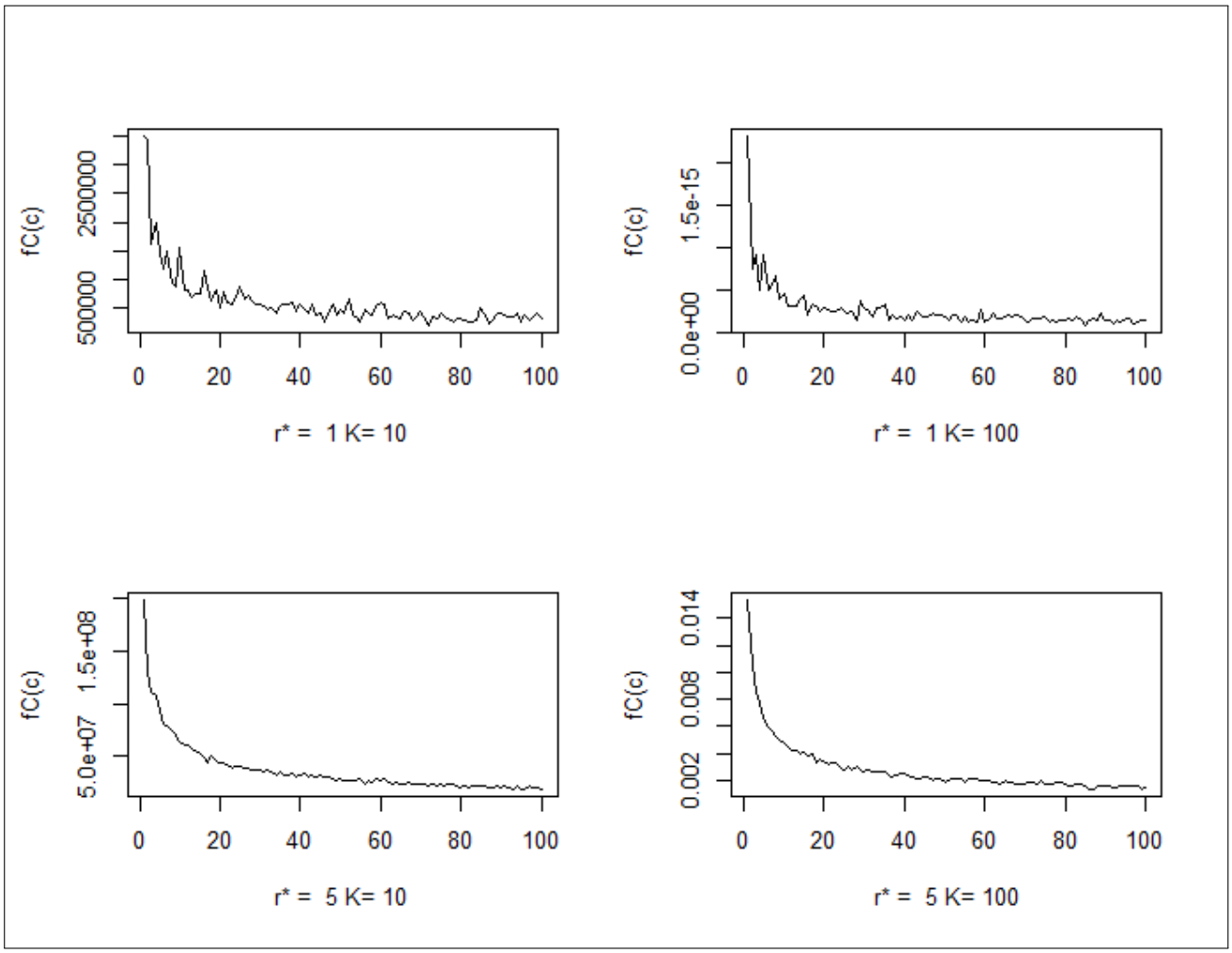

Figure **A-1**.: Plot function of $f_C(c)$ for $r^* = 1$, $r^* = 5$ and $K = 10$, $K = 100$.

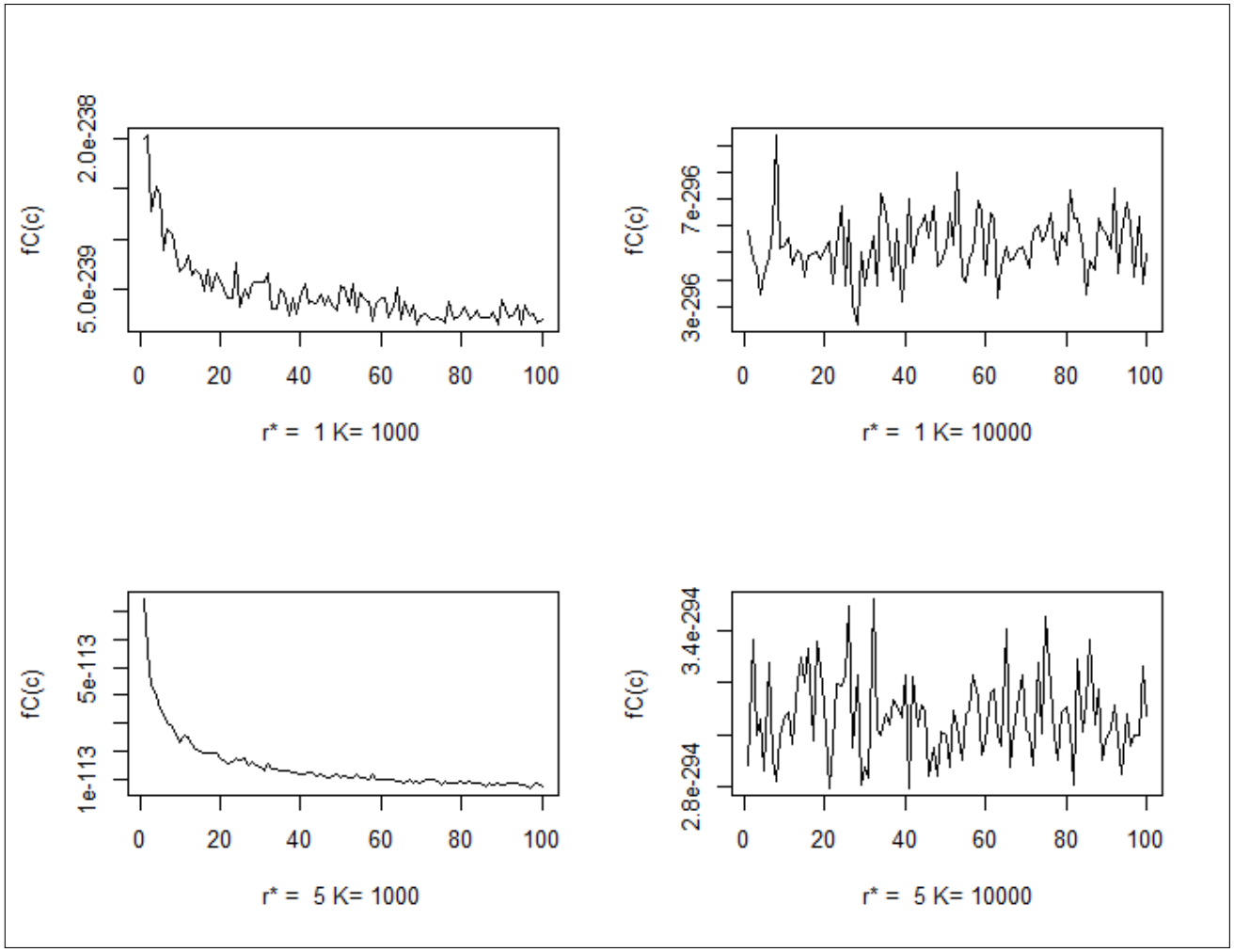

Figure **A-2**.: Plot function of $f_C(c)$ for $r^* = 1$, $r^* = 5$ and $K = 1000$, $K = 10000$.

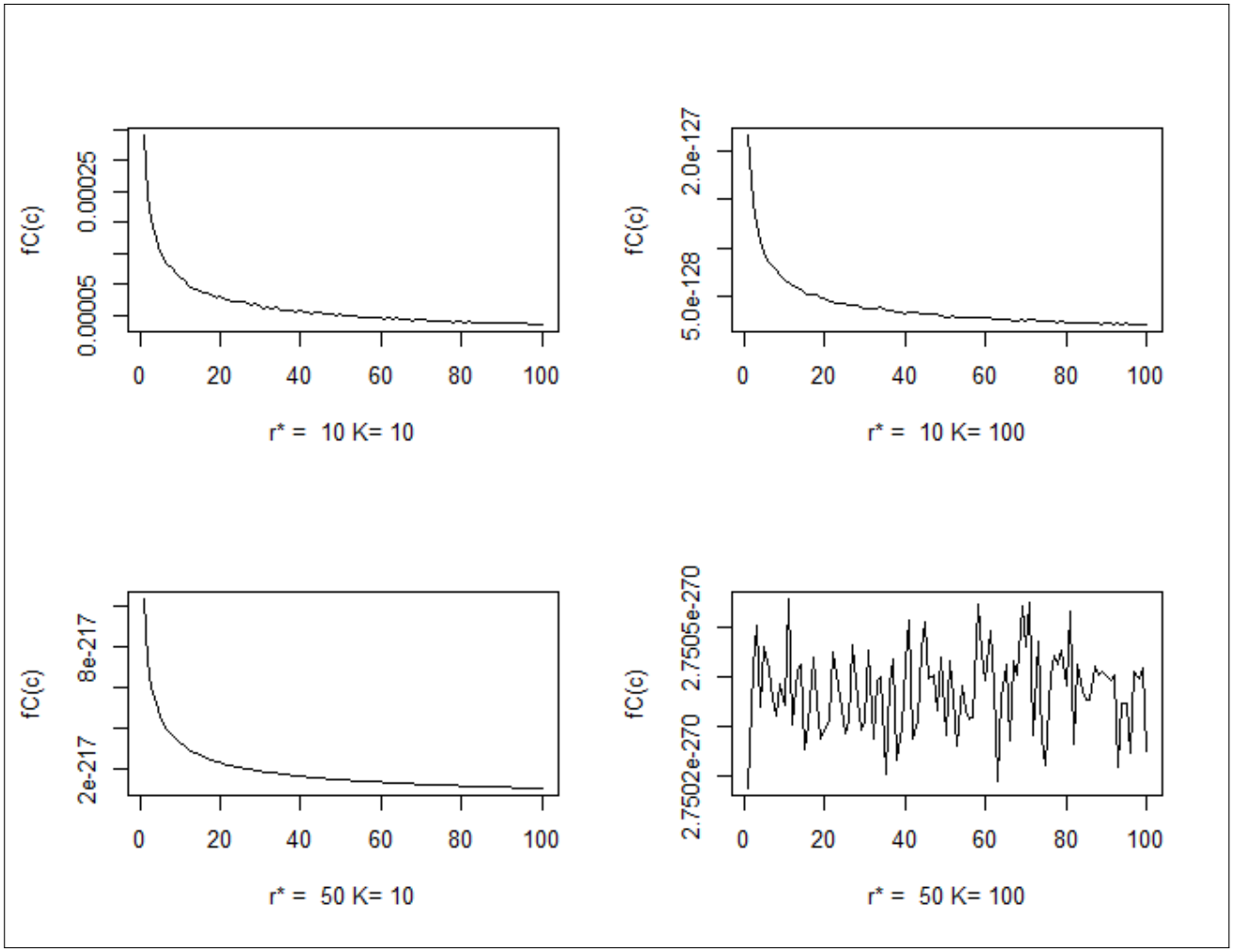

Figure **A-3**.: Plot function of $f_C(c)$ for $r^* = 10$, $r^* = 50$ and $K = 10$, $K = 100$.

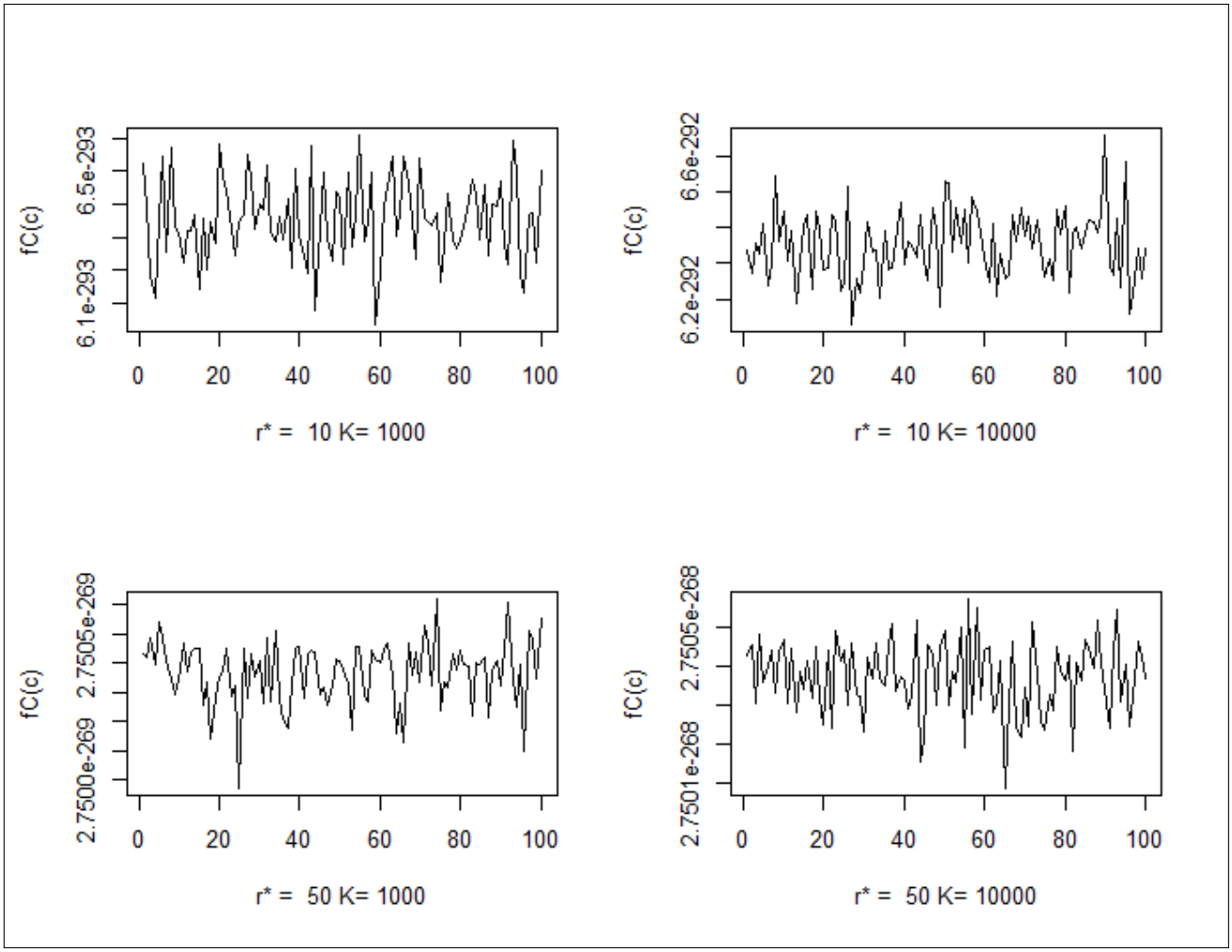

Figure **A-4**.: Plot function of $f_C(c)$ for $r^* = 10$, $r^* = 50$ and $K = 1000$, $K = 10000$.

B. Appendix: Behaviour of $L(r^*)$ as a function of X_i and K

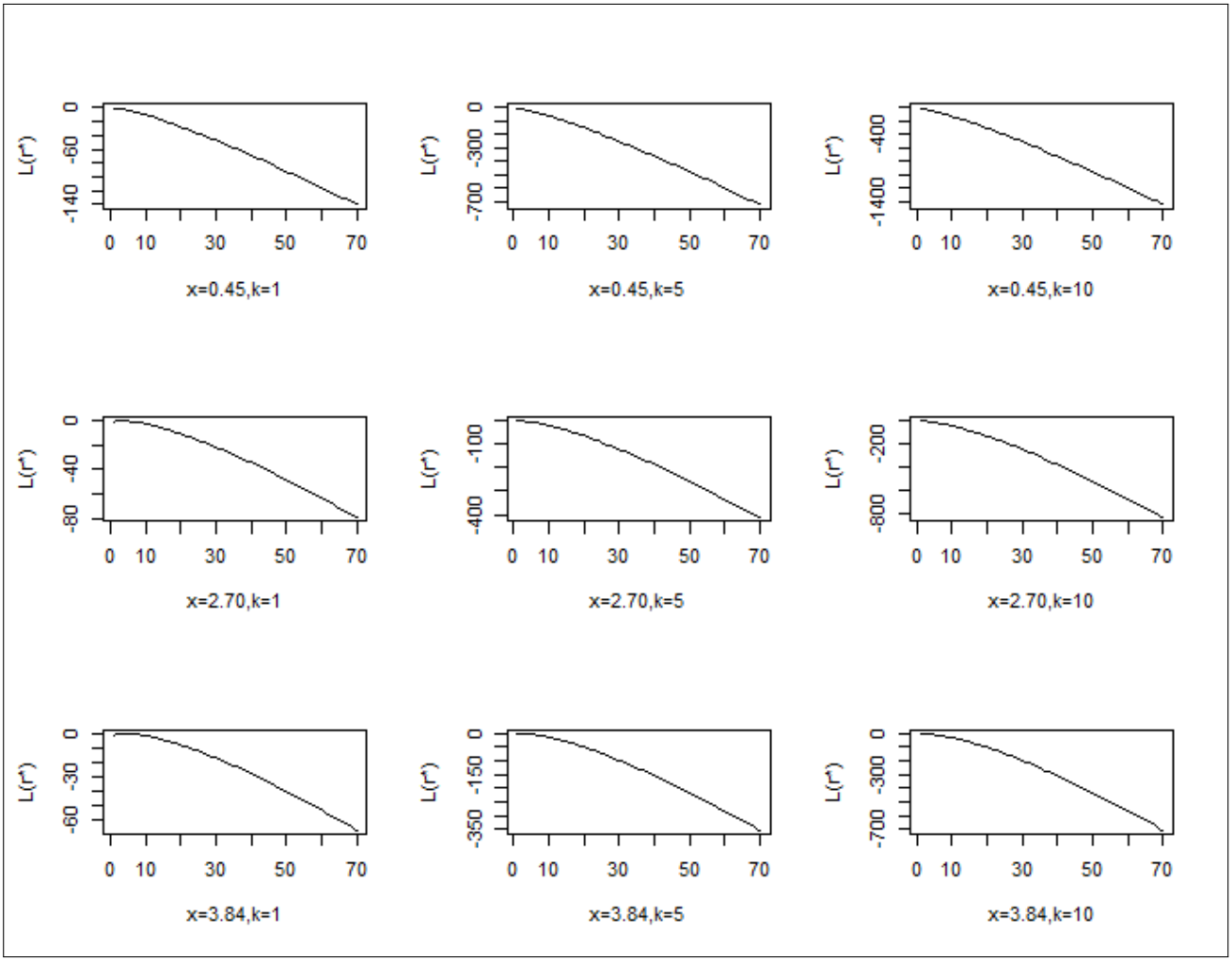

Figure **B-1**.: Behaviour of $L(r^*)$ for values of $X_i \leq 3.84$ and $K \leq 10$.

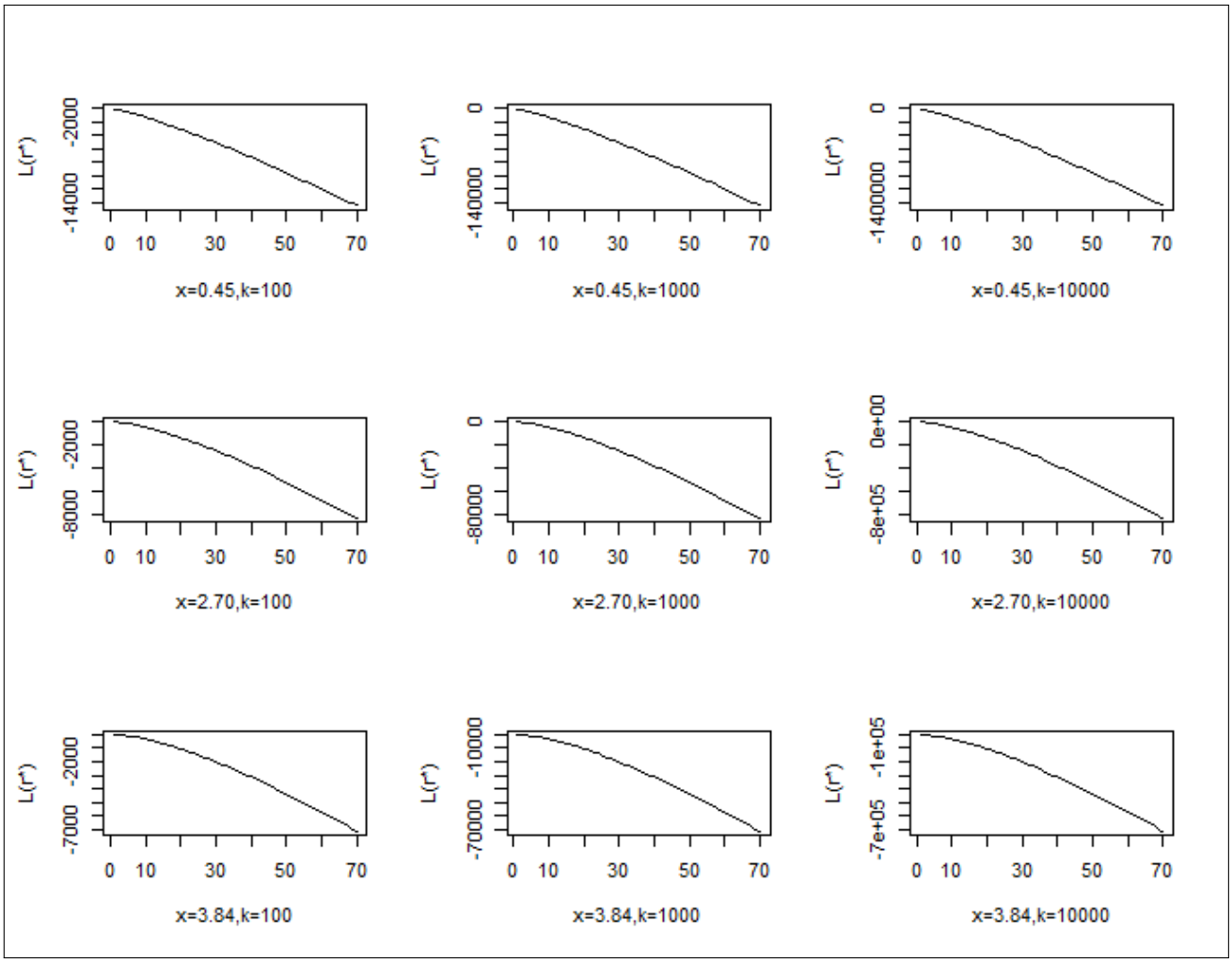

Figure **B-2**.: Behaviour of $L(r^*)$ for values of $X_i \leq 3.84$ and $100 \leq K \leq 10000$.

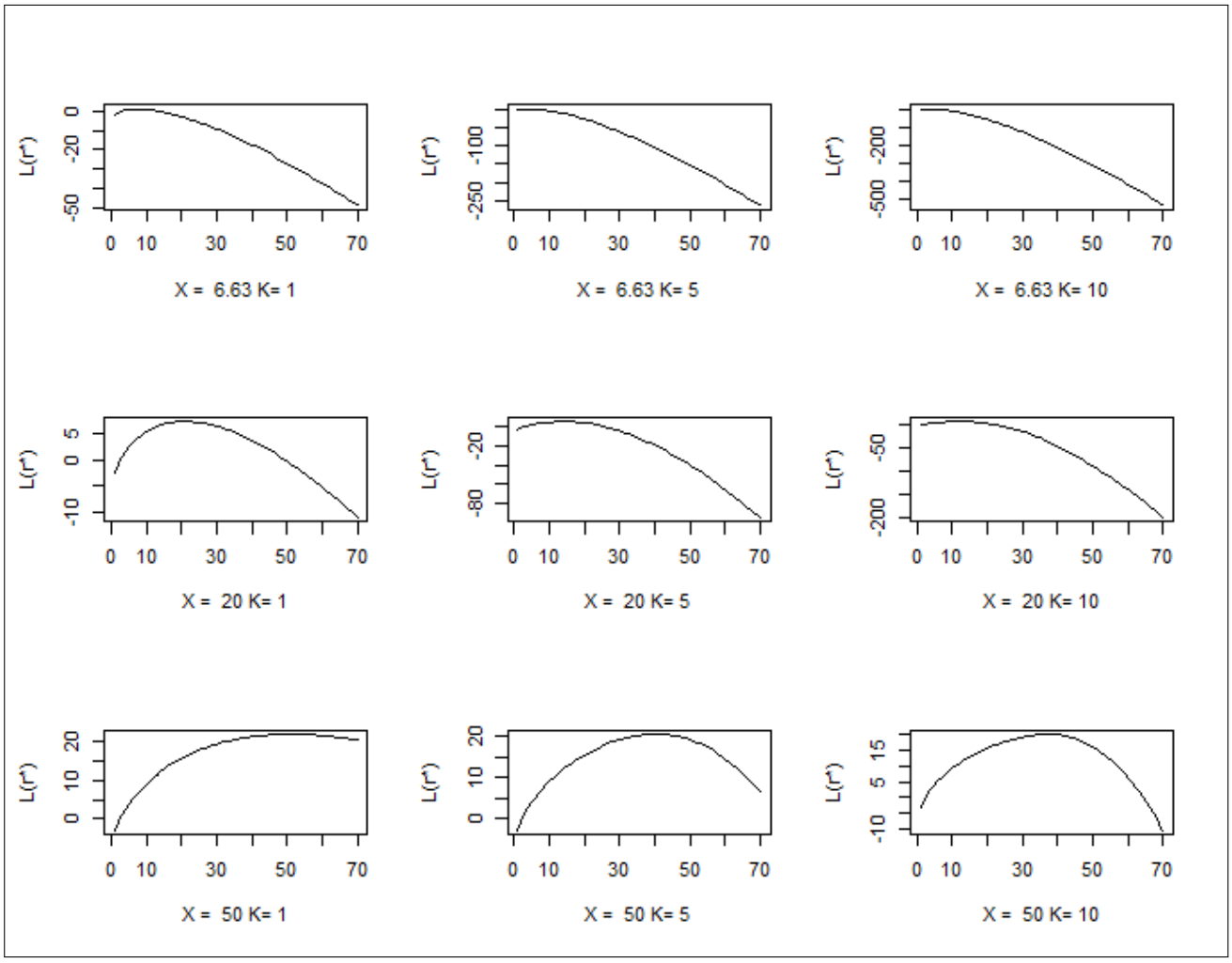

Figure **B-3**.: Behaviour of $L(r^*)$ for values of $6.63 \leq X_i \leq 50$ and $K \leq 10$.

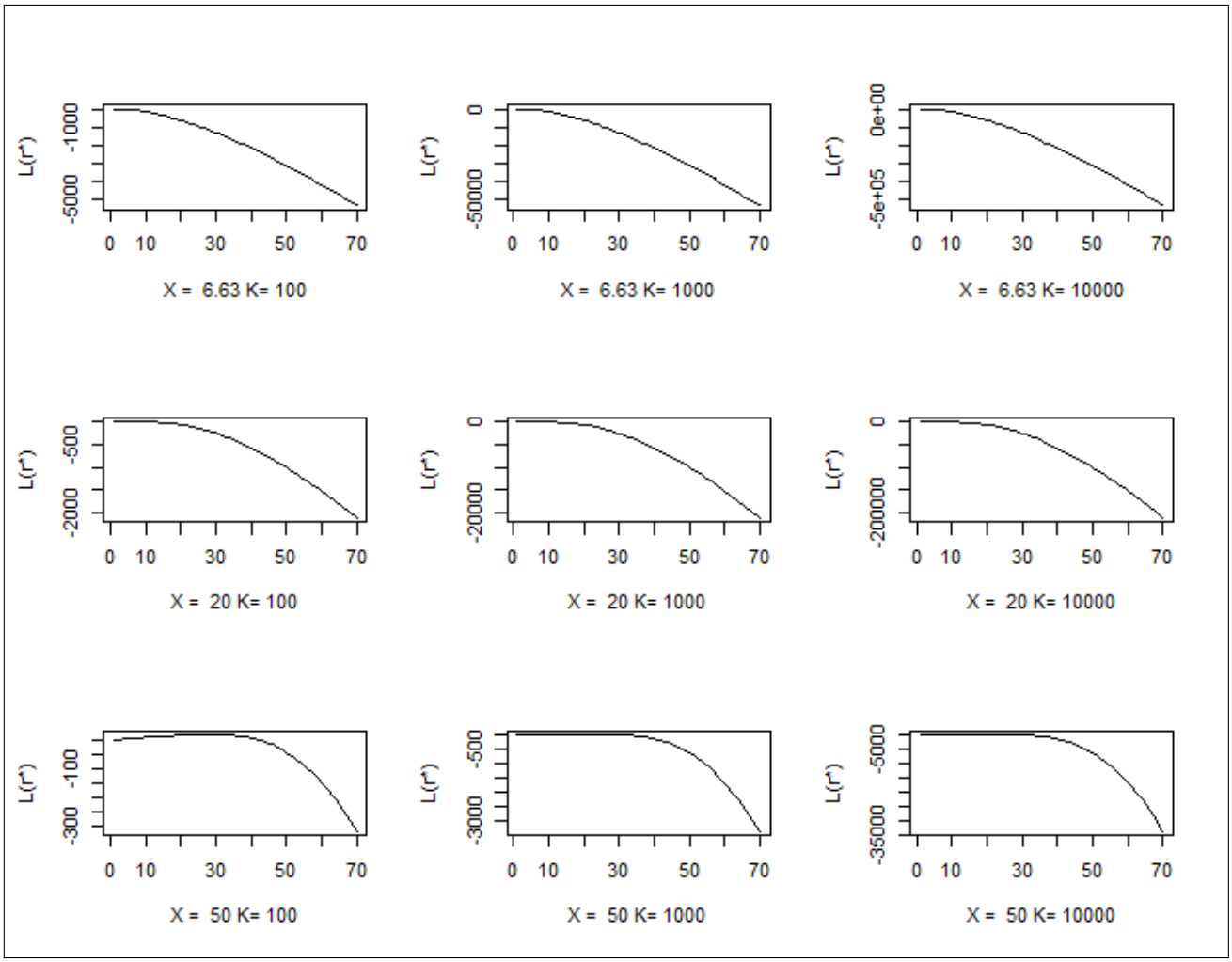

Figure **B-4**.: Behaviour of $L(r^*)$ for values of $6.63 \leq X_i \leq 50$ and $100 \leq K \leq 10000$.

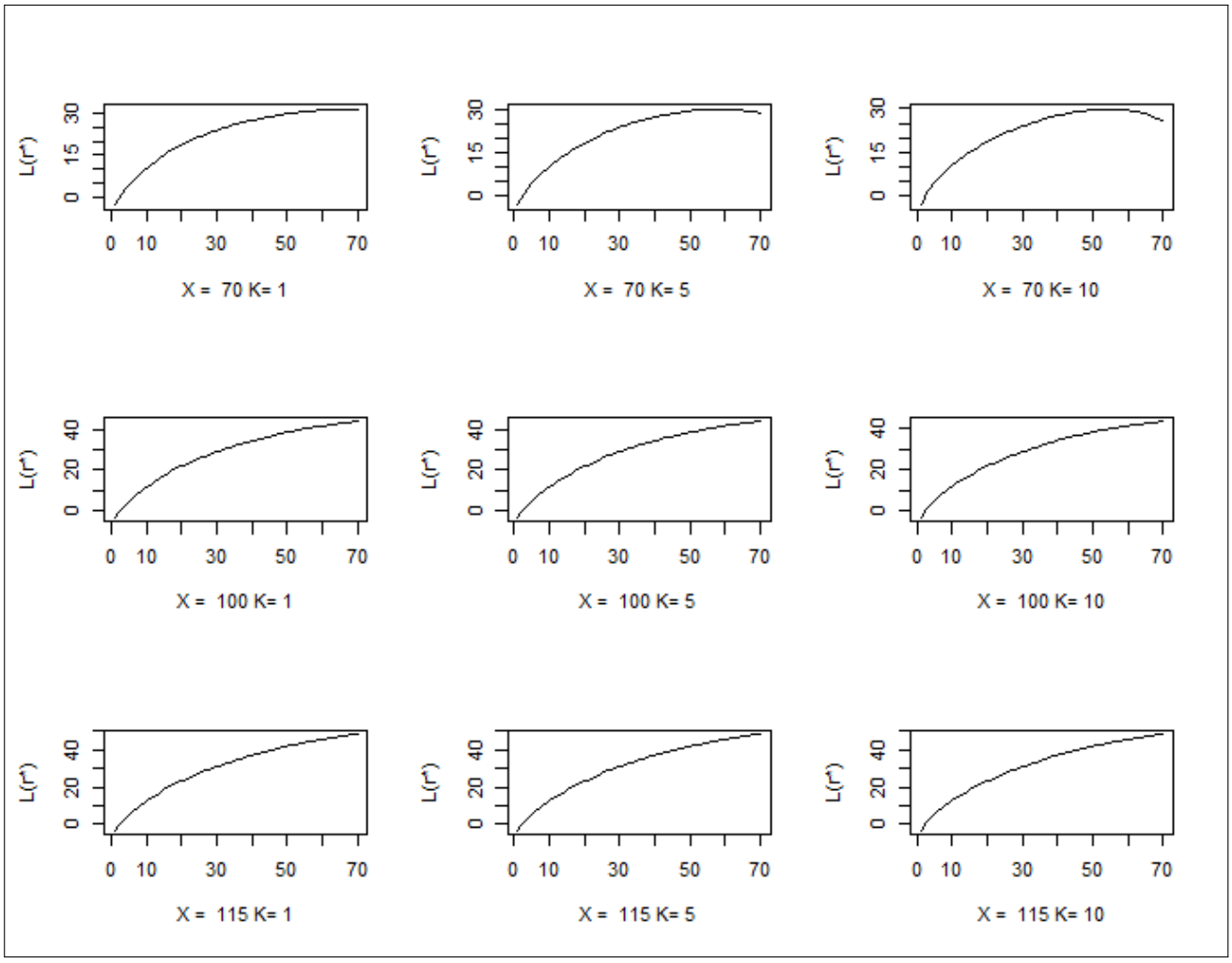

Figure **B-5**.: Behaviour of $L(r^*)$ for values of $70 \leq X_i \leq 115$ and $K \leq 10$.

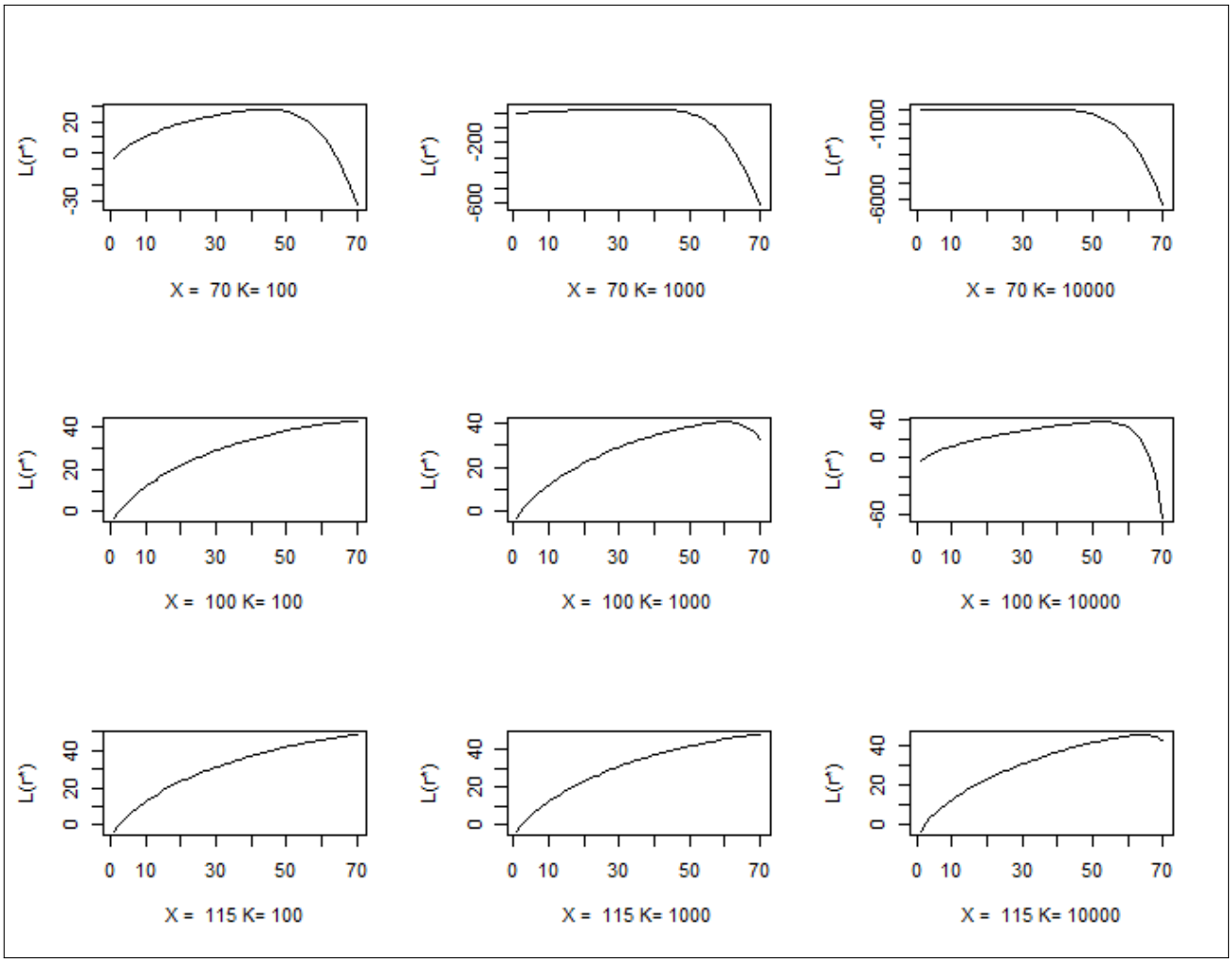

Figure **B-6**.: Behaviour of $L(r^*)$ for values of $70 \leq X_i \leq 115$ and $100 \leq K \leq 10000$.

C. Appendix: Programming in R

C.1. Estimation of r^* through the function optimize, calculation of raw and adjusted $p-$values, programming of the Quotient C and Manhattan plots in R

```
***********************************
REQUIRED LIBRARIES******************
***********************************

library(pracma) #Provides a large number of functions
                 as complete and incomplete gamma function
library(qqman) #Includes functions for creating Manhattan-plots

*******************************
OPTIMIZATION********************
*******************************

opt <- function(r,x,k){ #opt is the function in which "r" will be optimized
  g = as.numeric(gammainc(x = x/2,a = r/2)[1]) #Incomplete Gamma function,
                                               [1] means the limits of integration
                                               are upper= x/2 and lower= 0
  G = gamma(r/2) #Gamma function
  L = (k - 1)*log(g) + (((r/2) - 1)*log(x)) - (log(2)*(r/2)) - (k*log(G))
     #Function to be optimized (expression 5-8)
  return(L)
}

#FUNCTION gk(xk)(EXPRESSION 5-3). THIS FUNCTION INS NECESSARY FOR THE ESTIMATION
OF QUOTIENT C

g.k <- function(r,x,k){
  g = as.numeric(gammainc(x = x/2,a = r/2)[1]) #Incomplete Gamma function,
```

```
                                                    [1] means the limits of integration
                                                    are upper= x/2 and lower= 0
  G = gamma(r/2) #Gamma function
  gk = (k*((g/G)^(k - 1)))*(1/((2^(r/2))*(G)))*(x^((r/2) - 1))*exp(-x/2)
  return(gk)
}



**********************************************
Chi-SQUARED TEST AND P-VALUES*****************
**********************************************

SNP.summary <- function(D){ #D is a data frame
  K = ncol(D) - 1 #K is the number of SNPs in the dataser
  R <- matrix(data = 0,nrow = K,ncol = 10) #R is the object in which the results
                                           are saved
  R <- as.data.frame(R) #Output object type data frame
  colnames(R) <- c("SNPs", "X-squared","Ui(ki)","p-valor","p-valor Bonferroni",
              "p-valor BH", "p-valor Hochberg","p-valor Holm","r*","gk(xk)")
               #Names of the columns

 #LOOPS NECESSARY FOR THE CALCULATION OF STATISTICS

  for (i in 1:K) {
    l = i + 1 #The number of SNPs is counted
    t = chisq.test(x = as.numeric(D[,1]),y = as.numeric(D[,l])) #Chi-squared test

    #Statistic Ui(Ki) (expression 5-1)

    a = as.numeric(t$observed[4]) #SNP=1 (present), Y=1
    b = as.numeric(t$observed[3]) #SNP=1 (present), Y=0
    c = as.numeric(t$observed[2]) #SNP=0 (non present), Y=1
    d = as.numeric(t$observed[1]) #SNP=0 (non present), Y=0
    n = a + b + c + d

    R[i,1] <- colnames(D)[l] #SNP
    R[i,2] <- round(as.numeric(t$statistic),5) #Chi-squared valuesi
    R[i,3] <- (n*(((a*d) - (b*c))^2))/((a + b)*(a + c)*(c + d)*(b + d)) # Values of Ui(
    R[i,4] <- t$p.value #p-values for chi-squared test
    R[i,9] <- round(as.numeric(optimize(f = opt,lower = 1,upper = 100,maximum = T,
           x = R[i,3],k = K)[1])) #values of r* by the function optimize
    R[i,10] <- g.k(r = R[i,9],x = R[i,3],k = K) # The function g.k is estimated using
```

```
                                                    the values of r* of the column
                                                    9 and "x" (the  value of the
                                                    statistic Ui(ki))

  }
  #Adjusted p-values
  R[,5] <- p.adjust(p = R[,4],method = "bonferroni") #Adjusted p-value by
                                                       Bonferronis method
  R[,6] <- p.adjust(p = R[,4],method = "BH") #Adjusted p-value by BHs method
  R[,7] <- p.adjust(p = R[,4],method = "hochberg") #Adjusted p-value by
                                                     Hochbergs method
  R[,8] <- p.adjust(p = R[,4],method = "holm") #Adjusted p-value by  Holms method

  return(R)
}

#gral.SNP returns a table withe the number of SNPs associated and non associated
 according their raw and adjusted p-values

gral.SNP <- function(R){
  #R is a data frame with the function SNP.SUMMARY
  C <- matrix(data = 0,ncol = 6,nrow = 3)
  C <- as.data.frame(C)
  colnames(C) <- c("Crudo","Bonferroni","Holm","Hochberg","BH","C")
                 #Names of the columns
  row.names(C) <- c("Asociado","No asociado","Total")
                 #Names of the rows

  C[3,] = nrow(R) #Total number of SNPs

  C[1,1] = nrow(subset(x = R,subset = R$`p-valor` <= 0.05))
           #SNPs associated
  C[2,1] = nrow(subset(x = R,subset = R$`p-valor` > 0.05))
           #SNPs non associated

  C[1,2] = nrow(subset(x = R,subset = R$`p-valor Bonferroni` <= 0.05))
           #SNPs associated
  C[2,2] = nrow(subset(x = R,subset = R$`p-valor Bonferroni` > 0.05))
           #SNPs non associated

  C[1,3] = nrow(subset(x = R,subset = R$`p-valor Holm` <= 0.05))
           #SNPs associated
```

```
  C[2,3] = nrow(subset(x = R,subset = R$‘p-valor Holm‘ > 0.05))
           #SNPs non associated

  C[1,4] = nrow(subset(x = R,subset = R$‘p-valor Hochberg‘ <= 0.05))
           #SNPs associated
  C[2,4] = nrow(subset(x = R,subset = R$‘p-valor Hochberg‘ > 0.05))
          #SNPs non associated

  C[1,5] = nrow(subset(x = R,subset = R$‘p-valor BH‘ <= 0.05))
          #SNPs associated
  C[2,5] = nrow(subset(x = R,subset = R$‘p-valor BH‘ > 0.05))
          #SNPs non associated
  return(C)
}

#ASO.SNP is a function that identifies which SNPs are associated with at least one meth
Aso.SNP <- function(R){

  L <- subset(R,subset = R$‘p-valor Bonferroni‘ <= 0.05 | R$‘p-valor BH‘
    <= 0.05 | R$‘p-valor Hochberg‘ <= 0.05 | R$‘p-valor Holm‘ <= 0.05 )
  tab <- matrix(data = 0,nrow = nrow(L),ncol = 9)
  tab <- as.data.frame(tab)
  colnames(tab) <- c("SNP","Posicion","X","r*","Bonferroni","BH","Hochberg","Holm","C")

  tab[,1] <- L[,1] #SNP
  tab[,3] <- L[,3] #Value of statistic Ui(Ki), now denote as X
  tab[,4] <- L[,9] #r* value
  tab[,5] <- L[,5] #p-value by Bonferroni
  tab[,6] <- L[,6] #p-value by BH
  tab[,7] <- L[,7] #p-value by Hochberg
  tab[,8] <- L[,8] #p-value by Holm

  return(tab)

}

*********************************************
QUOTIENT C***********************************
*********************************************

stat.C <- function(data){
  #data: the dataset as in SNP.SUMMARY
```

```
  k <- as.numeric(ncol(data) - 1) #Number of SNPs
  C <- matrix(data = 0,nrow = k,ncol = 5)
  C <- as.data.frame(C)
  colnames(C) <- c("U","r*","gk","C","p_val")
  for (i in 1:k) {
    l = i + 1 #The number of SNPs is counted
    t = chisq.test(x = as.numeric(data[,1]),y = as.numeric(data[,l]))
      #Chi-squared test

    #Statistic Ui(Ki)
    a = as.numeric(t$observed[4]) #SNP=1 y P =1
    b = as.numeric(t$observed[3]) #SNP=1 y P =0
    c = as.numeric(t$observed[2]) #SNP=0 y P =1
    d = as.numeric(t$observed[1]) #SNP=0 y P =0
    n = nrow(data)
    C[i,1] <- (n*(((a*d) - (b*c))^2))/((a + b)*(a + c)*(c + d)*(b + d))
           #Statistic Ui(Ki)
    C[i,2] <- round(as.numeric(optimize(f = opt,lower = 1,
             upper = 100,maximum = T, x = C[i,1],k = k)[1]))
             #r* is calculated as in SNP.SUMMARY
    C[i,3] <- g.k(r = C[i,2],x = C[i,1],k = k) #g.k is calculated
    C[,5] = 0 #This column will save the p-values
  }
  for(i in 1:k){
    if(C[i,3] <= 0.000001){
      C[i,3] = 0.000001
    }
  }
  for (i in 1:k){
    C[i,4] <- (C[i,1])/(C[i,3]) #The Quotient C is calculated
  }
  return(C)
}

**********************************************
MANHATTAN PLOT*********************************
**********************************************

Manh.plot <- function(R,cromo,method){
    #cromo: a dataset of the chromosomal localization of each one SNP
  #method:  1-raw, 2-bonferroni,3-BH,4-hochberg, 5 holm
  if (method == 1 || method == 2 || method == 3 || method == 4 || method == 5) {
```

```
    m = 3 + method
    P <- R[,m] #p-values
    CHR <- cromo[,1]
    BP <- cromo[,2]
    SNP <- R$SNPs #Name of SNP
    X <- cbind(CHR,BP,P,SNP)
    X <- as.data.frame(X)
    colnames(X) <- c("CHR","BP","P","SNP")
    M <- manhattan(x = X,chr = "CHR",bp = "BP",
      p = "P",snp = "SNP",col = seq(2,27),cex = 0.6,
      cex.axis = 0.9, suggestiveline = F, genomewideline = F,ylim = c(0,10))
    abline(h = -log10(0.05),col = "red")
    return(M)
  }else{
    return("Error method")
  }
}
```

C.2. Probability density function of Quotient C

```
#Here, the probability density function of the Quotient C is calculated as well as
its graphic representation. The programming is given for r*=1.

***********************************
REQUIRED LIBRARIES******************
***********************************

library(pracma)

***********************************
EXPECTATION IN EXPRESSION (5-7)*****
***********************************

#A vector of Gamma values is created.

esperanza <- function(vec){
   E <- NULL #E is a object in which the results are saved
  for (i in 1:length(vec)) {
    e = as.numeric((gammainc(x = vec[i],a = r/2)[1])^(k - 1))
     For all the values in vec, the incomplete Gamma function is calculated
     if ( e <= max & e >= min) {
```

```
      E[i] = e
    }else if (e > max) {
      E[i] = max
    }else if (e < min) {
      E[i] = min
    }
  }
  return(mean(E)) #The return is the expected value of the object E
}

#To calculate the probability density function for r*=1, for example in
the dataset of hardness of the meat in sheep, the following values
are defined:

n = 500 #Amount of random Gamma values
k = 30776 #K, the number of SNPs in the dataset of hardness
           of the meat in sheep
r = 1 #r* value
v = (r + 1)/2 #The parameter of the Gamma function (expression (5-7))
cs = 100 #A amount of Quotient C

**********************************
PROBABILITY DENSITY FUNCTION********
**********************************

#The curve of the probability density function is created

fc <- NULL #The points of the curve are saved
for (i in 1:cs) {
  c <- as.numeric(i)
  xs <- rgamma(n = n,shape = v, scale = c/2) #Vector of Gamma random values
  xs <- (2*xs)/c
  xs <- as.vector(xs)
  esp <- esperanza(xs)
  num = (k*exp(-0.5)*(gamma(v))) #Numerator of the expression (5-7)
  den = (sqrt(2*pi*c)*(2^(r + 1))*(gamma(r/2)^k)) #Denominator of the
                                                   expression (5-7)

#Restriction over expression (5-7)
  if (den >= min & den <= max ) {
    den = den
  }else if (den < min) {
```

```
    den = min
  }else if (den > max ) {
    den = max
  }
  fc[i] = (esp)*(num/den) #Calculation of the expression (5-7)
}

************************************************
GRAPHS OF THE PROBABILITY DENSITY FUNCTION********
************************************************

cs = 100
n = 500
#Graph of expression (5-7) for values of r, k, cs y n
graphfc <- function(r,k){
  #r: r*
  #k: number of SNPs
  v = (r + 1)/2 #Parameter of the Gamma random variable
  fc <- NULL
  for (i in 1:cs) {
    c <- as.numeric(i)
    xs <- rgamma(n = n,shape = v, scale = c/2)
    xs <- (2*xs)/c
    xs <- as.vector(xs)
    esp <- esperanza(xs)
    num = (k*exp(-0.5)*(gamma(v)))
    den = (sqrt(2*pi*c)*(2^(r + 1))*(gamma(r/2)^k))
    if (den >= min & den <= max ) {
      den = den
    }else if (den < min) {
      den = min
    }else if (den > max ) {
      den = max
    }

    fc[i] = (esp)*(num/den)


  }
  plot(x = seq(1:cs),y = fc,type = "l",xlab = "",ylab = "fC(c)")
                       #Plot of the expression (5-7)
}
```

```
#allgraphfc graphs in a grid the probability density function of Quotient C
for the possible combinations of r* and  k given in l1 and l2.

allgraphfc <- function(l1,l2){
  for (i in 1:length(l1)) {
    for (j in 1:length(l2)) {
      graphfc(r = l1[i],k = l2[j])
      title(xlab = paste("r* = ",l1[i],"K=",l2[j]))
    }
  }
}

#A example is given below:

l1 = c(10,50) #r
l2 = c(100,1000) #k
dev.off()
par(mfrow = c(2,2))
allgraphfc(l1,l2)
```

C.3. Monte Carlo integration for $r^* = 1$

```
#The example below shows the procedure of estimation of the probability
density function of Quotient C by Mont Carlo integration when r*=1.

#A subset of SNPs with r*=1 is selected
s1 = subset(S,S$`r*` == 1) #S is a data frame of the function "stat.C"

#A histogram as well as the probability density function of Quotient C is shown
hist(s1$C,freq = F)
lines(density(s1$C))

**********************************
MONTE CARLO INTEGRATION*************
**********************************

D <- density(s1$C,n = 10000) #In a object called "D", 10000 points are saved,
                              they represent values of "x" and "y" according to
                              the probability density function of Quotient C

x <- D$x #"x" axis of the probability density function (expression (5-7))
y <- D$y #"y" axis of the probability density function (expression (5-7))
```

```
n = length(x) #Number of points
set.seed(13) #A seed in R
ys <- runif(n = n,min = 0,max = max(y)) #According to Monte Carlo integration,
                                n random points are created following a uniform
                                          distribution [0, maximum value of the
                                          probability density function].
                                          The loop transforms to
                                          zero those points that are above of
                                          the function and the points inside
                                          on the function are saved in a object
                                          called "cuenta".

cuenta <- NULL
for (i in 1:n) {
  if ( ys[i] < y[i]) {
    cuenta[i] = ys[i]
  }else{
    cuenta[i] = 0
  }
}

***********************************
MONTE CARLO INTEGRATION GRAPH*******
***********************************

plot(D,xlim = c(5000,max(x)),main = "",xlab = "c",ylab = "fC(c)")
par(new = T)
plot(x = x,y = cuenta,col = "red",pch = 20,xlab = "",ylab = "",axes = F,xlim = c(5000,m

#The "D" object is modified. Only the simulated points bigger than zero as well as the

cuenta <- cbind(x,cuenta)
cuenta <- subset(cuenta, cuenta[,2] > 0) #Simulated points bigger than 0
cuenta <- subset(cuenta, cuenta[,1] > 0) #"x" bigger than 0
tot = nrow(cuenta);tot #Total number of points inside of the function


***********************************
p-VALUE*****************************
***********************************

c = s1$C[1] #the value of "c" is defined
```

```
x0 <- subset(cuenta, cuenta[,1] >= c) #The values bigger or equal than "c"

nrow(x0)/tot #Proportion of points bigger than "c" (p-value)

#A loop is created. The procedure is repeatable for each one of the SNPs and the
p-values are saved.

for (i in 1:nrow(s1)) {
  c = s1$C[i]
  x0 <- subset(cuenta, cuenta[,1] > c)
  s1$p_val[i] = nrow(x0)/tot
}
```

C.4. Statistical power

```
#A connection to Bioconductor is necessary

source("http://bioconductor.org/biocLite.R")
#librerias necesarias para el calculo del FDR

**********************************
REQUIRED LIBRARIES*****************
**********************************

#The following libraries allow the estimation of the false discovery rate (FDR):

biocLite("multtest")
biocLite("qvalue")
library(multtest)
library(qvalue)

S <- stat.C(data)
s1 = subset(S,S$`r*` == 1)

#FDR estimation. Estimate the q-values for a given set of p-values. The q-value
of a test measures the proportion of false positives incurred (called the false
discovery rate) when that particular test is called significant

pv = sort(c(s1$p_val,rep(0,2606))) #A vector of all the p-values. In 2606 SNPs,
the p-value =0.
K = as.numeric(length(pv));K
fdr = qvalue(pv,0.05) #The q-value is calculated
```

```
plot(fdr)
aso = as.numeric(table(pv <= 0.05)[2]);aso #Amount of SNPs with p-values<=0.05
pro = rep(pbinom(0,K,aso/K,lower.tail = F),K) #The probability that the number
                                                of rejected test>0 (See section
                                                5.4 Statistical power).

pot = (pro - fdr$qvalues) #The statistical power is calculated

#Considerations over expression (5-11)
for (i in 1:length(pot)) {
  if (pot[i] < 0) {
    pot[i] = 0
  }
}
plot(pot,type = "l",ylab = "Statistical Power",xlab = "K",main = "Quotient C",pch = "."
                #Line graph of statistical power
```