BACHELOR'S DEGREE THESIS

Degree in Mathematics

Title: A genetic association study of colon cancer

Author: Júlia Densalat Ventayol

Advisor: Jan Graffelman

Department: Department of Statistics and Operations Research

Academic year: 2013/2014

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Bachelor's degree thesis

A genetic association study of colon cancer

Júlia Densalat Ventayol

Advisor: Jan Graffelman

Statistics and Operations Research

To my family, who always had faith in me.

Abstract

Keywords: Statistics, genetics, colon cancer, single nucleotide polymorphism (SNP).

The project of this bachelor degree concerns the statistical analysis of a genetic database provided by Professor Victor Moreno, epidemiologist of the Hospital de Bellvitge in Barcelona. The goal is to investigate associations between a particular area of the genome and colon cancer. To achieve the objective, some basic statistical genetic models such as the codominant test, the exact test, the allele test and the Cochran-Armitage test will used to analize the data. These tests reveal a few clusters of genetic markers that appear to be associated with colon cancer.
All statistical analysis, computations and graphics are made with the statistical enviroment R.

Contents

Introduction

The project of this bachelor degree concerns the statistical analysis of a genetic database provided by Professor Victor Moreno, epidemiologist of the Hospital de Bellvitge in Barcelona.
The database contains genetic information for both healthy people and for people suffering from colon cancer.

We present a genetic association analysis that will enable us to investigate associations between genetic markers and a disease trait, in this case, colon cancer.

Genetic association studies are performed to determine whether a genetic variant is associated with a disease or trait: if association is present, a particular allele or genotype of a polymorphism or polymorphisms will be seen more often than expected by chance in an individual carrying the trait.

There exist two types of genetic association studies: *Family-based association* studies and *Population-based association* studies.

Family-based association studies

Family based association studies aim to avoid the potential confounding effects of population stratification by using the parents or using unaffected siblings as controls for the case, which is their affected offspring/ siblings. If an allele increases the risk of having a disease then that allele is expected to be transmitted from parent to offspring more often in populations with the disease.

Population-based association studies

Population based association studies work with samples of individuals without family relationship that are therefore considered to be independent individuals. Statistical techniques used for population-based association studies, like for instance logistic regression, typically assume independent observations. A typical design for population-based studies is a case-control design, with a random sample of cases (individuals affected with the disease) and a random sample of controls.

We can also describe an association study by saying whether it is a *Genome wide association study* or a *Candidate region study.*

Genome-wide association studies (GWAS)

A genome-wide association study is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait. It scans the entire genome for common genetic variation.

Candidate region studies

The candidate gene approach to conducting genetic association studies focuses on associations between genetic variation within pre-specified genes of interest and phenotypes or disease states.

The genetic association study that we present is a Population-based association study and a Candidate region study focused on chromosome 6. The chromosome 6 has been chosen because epidemiologists suspect that a region of this chromosome is likely to present associations between colon cancer and the markers located on it.

A case-control design is used to do the study, in which a series of cases affected with the disease of interest (colon cancer) are collected together with a series of control individuals.

1 Statistical and molecular genetics

In this chapter we present some basic genetic terminology that is necessary to understand the rest of this project.

Statistical Genetics deals with the analysis of genetic data and inherited traits. Genetic data refers to the biological material that is inherited during reproduction via egg and sperm cells.

The *Human genome* refers to all of the basic biological material that is transmitted from parents to offspring.
The heritable material is stored in *chromosomes* in the nucleus of every cell. Each chromosome is composed of long strands of DeoxyriboNucleic Acid (DNA). The DNA determines how proteins are manufactured in the body and is the basic biological material of inheritance. The human genome consists of 23 pairs of chromosomes.
Genes are largely contiguous stretches of DNA.
Polymorphic means the data at a particular location in a chromosome can have more than one possible variant. A *polymorphism* is a polymorphic genetic locus.
The different variants at a locus are called *alleles*. When there are only two variants, we refer to them as $'A'$ and $'a'$. An individual with two copies of A (one on each of the two chromosomes) is called *homozygous* AA. An individual with one A and one a is called *heterozygous* or Aa. Finally an individual with two copies of a is *homozygous* aa. The *genotype* of an individual refers to the pair of alleles at a location.
Recall that each person has two copies of each autosomal chromosome, so at any specific locus, each person has two alleles, one inherited from each parent. The *Genotype frequency* in a population is the number of individuals with a given genotype divided by the total number of individuals in population, so it is the proportion of a given genotype in a population.
The *Allele frequency* in a population is defined as the proportion of chromosomes carrying that allele, regardless of the pairing within individuals. Suppose that we have a sample of size n from a population with a proportion, p, of A alleles. Then to estimate p, we simply count the number of chromosomes carrying the A allele and divide by $2n$, the number of chromosomes.
The *Minor allele frequency (MAF)* refers to the frequency at which the least common allele occurs in a population.

A *trait*, also known as the *phenotype*, is used to mean individual characteristics that have a heritable basis.
A *genetic marker* is a gene or DNA sequence with a known location on a chromosome that allows us to distinguish genetic differences in individuals. Having marker data for samples of families has enabled to create a variety of methods to find the chromosomal location of a disease-causing gene.
A *disease susceptibility locus* indicates a specific genetic locus, which has a variant associated with a disease.
The simplest type of genetic marker is a single nucleotide polymorphism (SNP). A SNP is the

most common type of genetic variation among people. It is a variation at a single position in a DNA sequence among individuals. Recall that the DNA sequence is formed from a chain of four nucleotide bases: A, C, G, and T. If more than 1% of a population does not carry the same nucleotide at a specific position in the DNA sequence, then this variation can be classified as a SNP.

Although a particular SNP may not cause a disorder, some $SNPs$ are associated with certain diseases. These associations allow scientists to look for $SNPs$ in order to evaluate an individual's genetic predisposition to develop a disease. In addition, if certain $SNPs$ are known to be associated with a trait, then scientists may examine stretches of DNA near these $SNPs$ in an attempt to identify the gene or genes responsible for the trait.

Over 10 million $SNPs$ have been identified in the human genome. $SNPs$ occur at a rate of about 1 in 300 nucleotides.

From a statistical point of view a SNP is a categorical variable. Theoretically, this categorical variable can have 10 possible categories if the SNP has all four possible alleles (A,C,G,T). In practice almost all $SNPs$ are bi-allelic, and therefore only 3 genotypes will exist. In this case, a SNP is a categorical variable with 3 classes.

A *genetic model* describes the relationship between an individual's genotype and their phenotype.

2 Describing the data

In this chapter we present the data set being studied and give a descriptive analysis of the data.

2.1 Structure of the database

The dataset we are using to do this study contains two objects, one called "*calls*" and the other called "*annot*".

Object "*calls*" is a table with 1685 rows and 146 columns. Each row represents a bi-allelic polymorphism and each column represents an individual. Our study is based, then, on information about 146 individuals.
There's also an index indicating whether an individual is sick of colon cancer or not. The index has value M if the individual has the disease and N if not. Table "*calls*" contains numbers from 1 to 3. Number 1 indicates homozygous AA, number 2 indicates heterozygous Aa and number 3 indicates homozygous aa. Further, we can occasionally find 'NA' which means the value is missing. This table then, gives us the genotypes of each individual for each marker.
To start with, we will transpose the table "*calls*" in order to have the individuals represented in rows (we will use the same name for this table).
Thus, $calls_{ij} = 1$ means the individual i is homozygous AA for the marker j.
The genotypes of the first 10 individuals for the 10 first markers are given below.

	1	2	3	4	5	6	7	8	9	10
1_N	1	1	3	3	1	3	1	1	1	1
2_N	1	2	2	1	3	1	2	2	3	1
3_N	3	1	3	NA	1	3	3	3	1	1
4_N	1	2	2	2	2	2	1	1	3	1
5_M	2	1	3	2	2	2	3	3	2	1
6_N	1	2	2	2	2	2	1	1	3	1
7_N	1	1	3	2	2	2	2	2	1	1
8_N	3	1	3	3	1	3	3	3	1	1
9_N	2	1	3	3	1	3	2	2	1	1
10_M	2	1	3	2	2	2	3	3	1	1

Table 2.1: Genotypes of the first 10 individuals for the first 10 markers

The object "*annot*" contains information of the markers. It is a table of 1685 rows. Each row represents a marker and for each one of them we have the two different variants (alleles). We also have the minor allele frequency and also the position on chromosome 6 of each marker measured

in base pairs.
Let's see all this information for the first 10 markers.

	Allele.A	Allele.a	Minor.Allele.Caucassian	MAF	position
1	G	T	T	0.28	1
2	A	T	T	0.26	338
3	G	T	G	0.26	4815
4	A	C	A	0.38	4864
5	A	G	G	0.40	4876
6	A	G	A	0.38	4939
7	C	T	T	0.48	17064
8	C	T	T	0.48	17279
9	C	T	T	0.42	19412
10	C	T	T	0.04	23888

Table 2.2: Information of the first 10 markers from objet annot

For example, the first row indicates that the first bi-allelic polymorphism marker on chromosome 6 is the one that occupies the first position on the chromosome. It also indicates that the two variants that this marker can take are G and T and the Minor Allele Frequency (allele T) is 0.28833.

We have the genotypes of 1685 markers of 146 individuals. To start with, let's find out how many of them are controls (don't have the disease) and how many are cases (have colon cancer).

	# individuals
Cases	47
Controls	99

Table 2.3: Composition of the database

As we see in the table, we have 47 cases and 99 controls.

2.2 Missing data

What is next is to investigate whether the dataset we are working with is good enough. That means it doesn't have too many missing values.

We may want to know, for each individual, the % of missing genotypes. We do this in order to decide if all the individuals are useful. For example, if one individual had missing genotypes for more than the 50% of the markers, then we would remove this individual from the study.
In this case, we will consider a 10% as a limit.

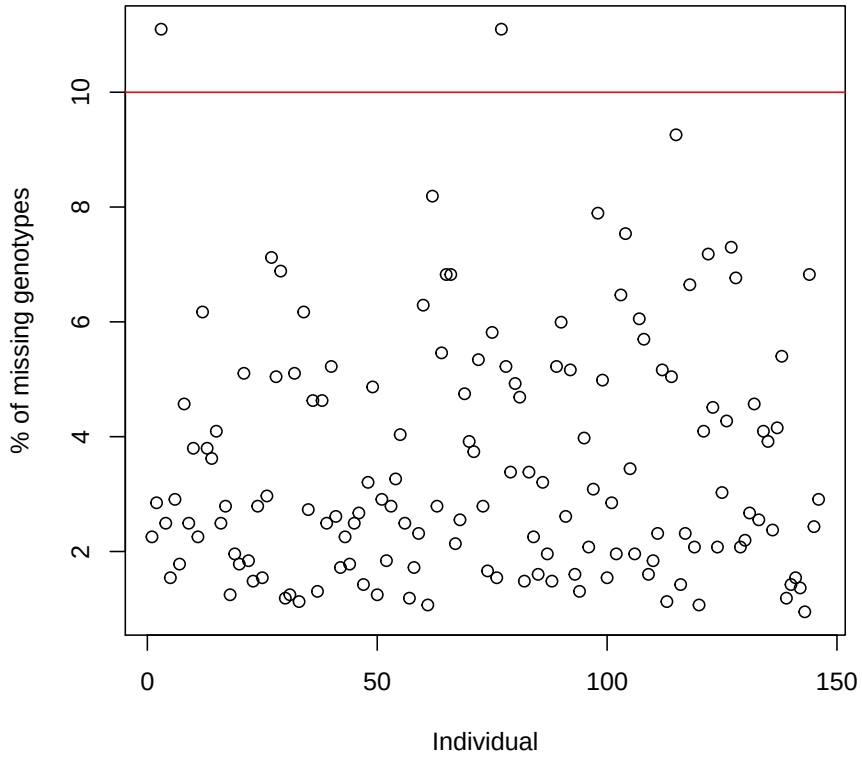

Figure 2.1: Plot of the percentage of missing values for each individual of the database

Figure 2.1 shows that almost all individuals have less than 10% missing values, and that there is no need to remove individuals from the database.

The next step is to calculate for each marker, the % of individuals that don't have a genotype of the marker. We do this in order to detect if there is a marker that we can't use to do the study. For example, if for the 50% of the individuals, we don't have the genotype of a particular marker, then we do not have enough information to analyze this marker and so we will remove it from the database.

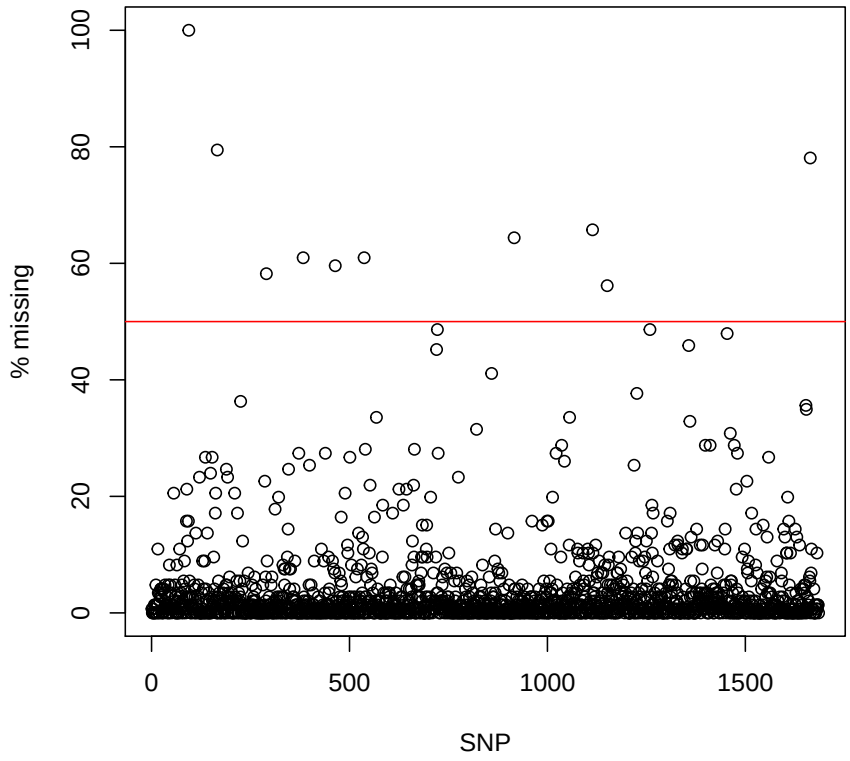

Figure 2.2: Plot of the percentage of missing values for each SNP of the database

Figure 2.2 shows that there are a few markers with a lot of missing values. Markers with more than 50% missing values were considered unreliable and were removed from the database. Initially we had 1685 markers but after this removal we got 1675.

2.3 Distribution of Minor Allele Frequency

Although, object "*annot*" contains the Minor Allele Frequency (MAF) for each marker, we are going to calculate it from the marker data that we have and we are going to estimate its probability function by making an histogram.

Calculation of Estimated Allele Frequencies from a Sample of n Subjects

Given the genotype counts for the sample:

$$\begin{aligned} n_{AA} &= \text{number out of } n \text{ with genotype AA} \\ n_{Aa} &= \text{number out of } n \text{ with genotype Aa} \\ n_{aa} &= \text{number out of } n \text{ with genotype aa} \end{aligned}$$

where $n_{AA} + n_{Aa} + n_{aa} = n$. The sample proportion of A alleles,

$$\hat{p} = \frac{(2n_{AA} + n_{Aa})}{2n} \tag{2.1}$$

estimates the A allele frequency. With a two allele system, the a allele frequency is $\hat{q} = 1 - \hat{p}$.

Thus, $MAF = min(\hat{p}, \hat{q})$.

After calculating the MAF for each SNP of the database we make a histogram with the results that have been obtained.

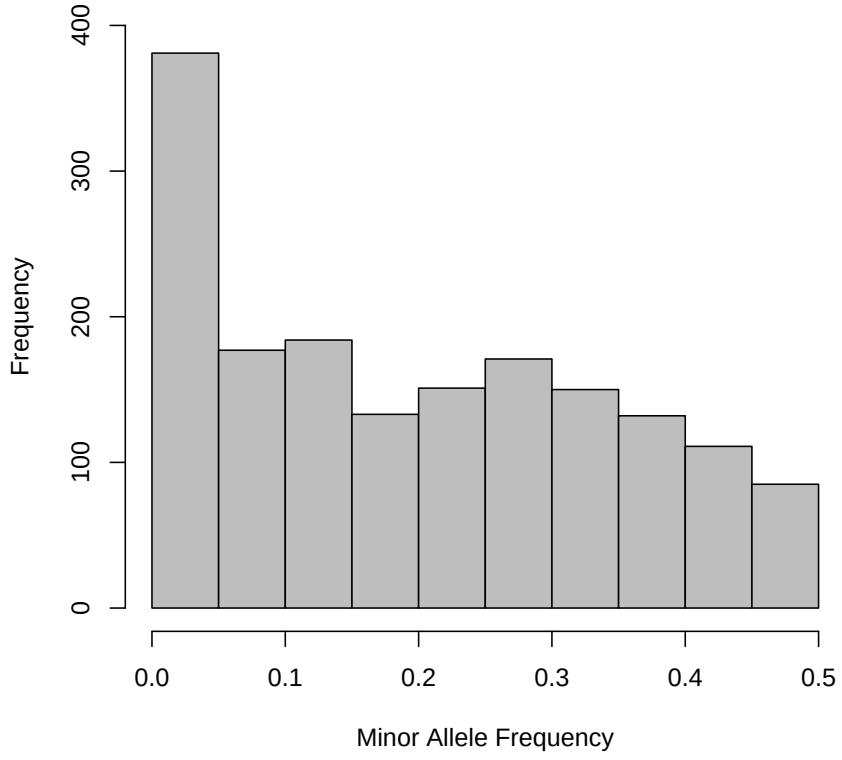

Figure 2.3: Histogram of the Minor Allele Frequency

Figure 2.3 shows that the Minor Allele Frequency is not uniformly distributed, but that markers with a low MAF (in the range $0.0 - 0.1$) are relatively more common. Marker with a 0.1 or low MAF are generally considered less informative because they have a lower variance.

How many monomorphic SNPs do we have?

A monomorphic SNP is a SNP for which a single form or allele can be identified in the population of interest. Equivalently, a SNP is monomorphic if $MAF = 0.00$.

As it is shown in figure 2.3, there are a lot of markers with a MAF in the range $0.00 - 0.05$ and so there can be markers with $MAF = 0.00$. We are going to report the number of monomorphic markers and eliminate them.

Doing the appropriate calculations we find out that 109 markers of our data marker are monomorphic. That means a single genotype is found for these markers in our population, and so these markers have zero variance and are not informative for our study. After the removal of these markers we now dispose of 1566 markers.

We now repeat the histogram of the MAF. This time we will not take into account the monomorphic SNPs.

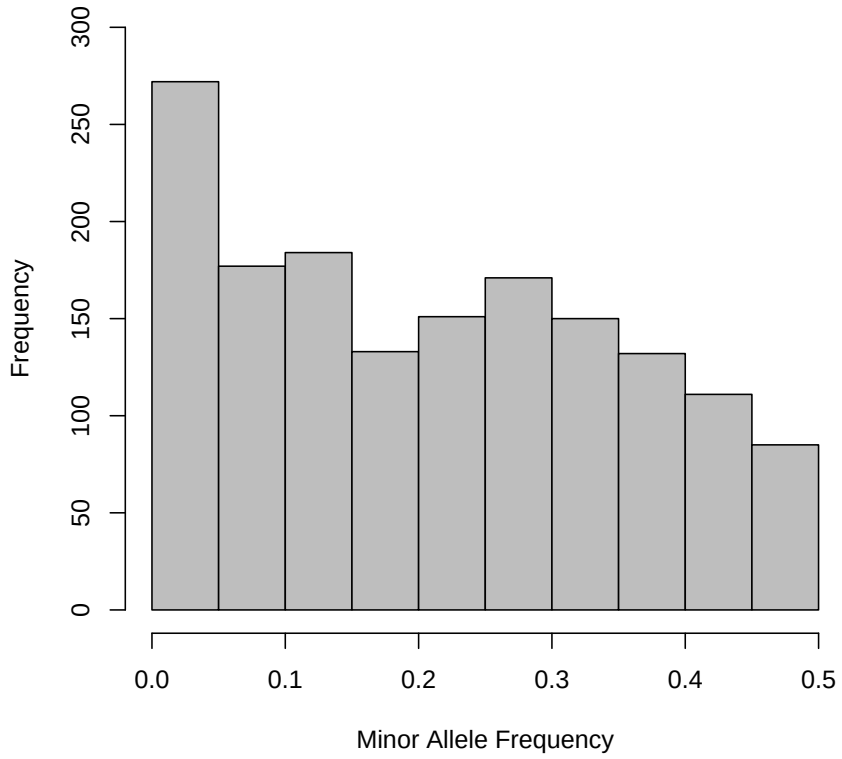

Figure 2.4: Histogram of the Minor Allele Frequency of the non-monomorphic SNPs

Comparing 2.4 with 2.3 we see that before removing the monomorphic SNPs we had nearly 400 markers in the range 0.0 - 0.05, but now we have less than 300. In any case, the MAF is not uniformly distributed.

2.4 Hardy-Weinberg equilibrium

In 1908, Godfrey Hardy [1] and Wilhelm Weinberg [2] independently derived a formula relating allele frequencies in parents to genotype frequencies in offspring.

Hardy-Weinberg equilibrium principle (HWE)

Let p be the frequency of the A allele and q be the frequency of the a allele, such that $p + q = 1$. A population is in Hardy-Weinberg Equilibrium if the genotype frequencies in the offspring are given by:

[1] Godfrey Harold Hardy (7 February 1877 - 1 December 1947 was an English mathematician, known for his achievements in number theory and mathematical analysis

[2] Dr. Wilhelm Weinberg (Stuttgart, December 25, 1862 - Tübingen, November 27, 1937) was a German physician credited as the first to explain the effect of ascertainment bias on observations in genetics

$$f_{AA} = P(AA) = p^2$$
$$f_{Aa} = P(Aa) = 2pq$$
$$f_{aa} = P(aa) = q^2$$

If this happens, then the frequency of the A allele among the offspring is also p and that means that under Equilibrium, allele frequencies will not change from generation to generation. Consequently, the genotype frequencies will also remain the same over the generations.

An alternative formulation of HWE is obtained by squaring the heterozygote frequency giving:

$$f_{Aa}^2 = 4p^2q^2 = 4f_{AA}f_{aa}$$

which gives an exact mathematical relationship between the 3 genotype frequencies.

There are some assumptions required for this formula to hold:

- The organism we are studying is diploid
- There is sexual reproduction
- Mutation can be ignored
- Migration is negligible
- Natural selection does not affect the trait under study
- Random mating

Despite none of these assumptions is likely to hold exactly in any populations, the principle provides a good approximation for population genotype frequencies.

Genetic markers are, in general, expected to follow HWE. If they do not follow the equilibrium is probably because of genotyping error. Genotyping error consists of confounding homozygotes with heterozygotes. E.g. a heterozygote AB is classified in the laboratory as AA or the reverse.
For this reason, markers need to be checked for HWE as part of a quality control procedure.

Do our markers follow the HWE?

To do this check for our markers, we are going to use an R package called $HardyWeinberg$ [1]. It constructs ternary plots for genotypic compositions for bi-allelic marker data.
The most important routine of the package is the $HWTernaryPlot$ function that draws a ternary plot for three-way genotypic compositions (AA, Aa, aa), and represents the acceptance region for different tests for HWE in the plot.

HWE can be inferred from the position of the marker in the ternary plot (see Figure 2.5). A marker that is in perfect equilibrium is on the parabola inside the ternary diagram. Markers that are, statistically speaking, in equilibrium, will be close to the parabola, inside the banana-shaped acceptance region in the plot (Graffelman & Morales) [2].

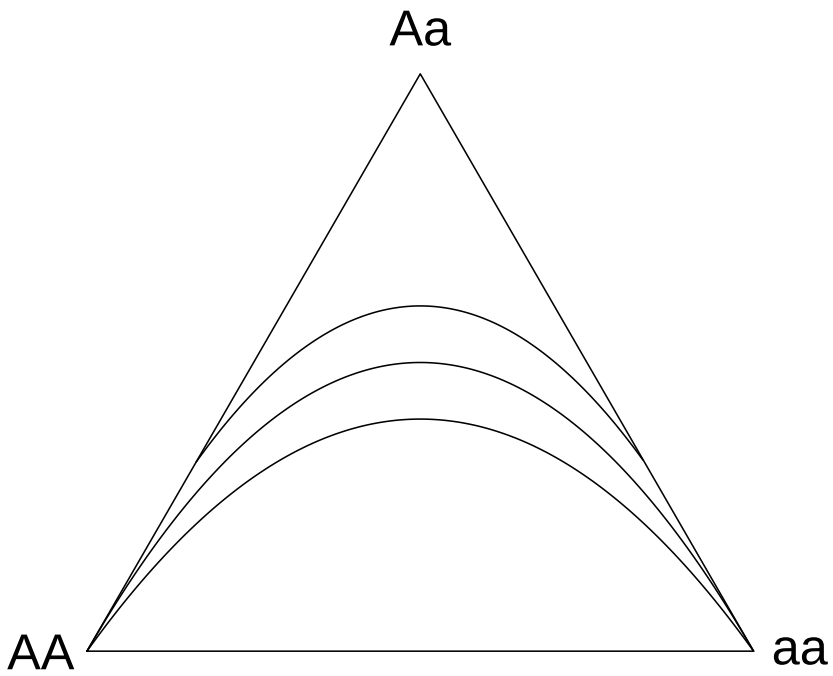

Figure 2.5: Ternary plot showing the parabola and the acceptance region corresponding to a Chi-square test

Different statistical tests for HWE can be done graphically with this routine: the ordinary chi-square test, the chi-square test with continuity correction and Haldane's exact test.

How do we obtain expressions for the curves that bound the acceptance region for HWE, for example, for the χ^2 test?

Let f_{AA}, f_{Aa} and f_{aa} be the genotype frequencies, and e_{AA}, e_{Aa}, e_{aa} the expected frequencies under HWE. Let n be the sample size and D the deviation from independence for the heterozygote given by $\frac{1}{2}(f_{Aa} - e_{Aa})$. Then the χ^2 statistic for testing HWE is:

$$X^2 = \frac{D^2}{p^2 q^2 n} \tag{2.2}$$

Where $X^2 \sim \chi_1^2$

By substituting $D = \frac{1}{2}(f_{Aa} - e_{Aa})$, we can rewrite Eq.(2) and express the relative sample frequency of heterozygotes r_{Aa} in terms of the allele frequency p to obtain the parabolas:

$$r_{Aa} = 2pq \pm 2pq\sqrt{\frac{X^2}{n}} \tag{2.3}$$

Sample heterozygote frequencies lie on the upper parabola if $D > 0$ which means heterozygote excess and on the lower parabola if $D < 0$ which means heterozygote dearth. Exact HWE is achieved when $X^2 = 0$.

For $D > 0$, we reject HWE when $X^2 > \chi_1^2(\alpha)$ [3] or equivalently when r_{Aa} exceeds $2pq + 2pq\sqrt{\frac{\chi_1^2(\alpha)}{n}}$.

For $D < 0$, we reject HWE when $X^2 < \chi_1^2(\alpha)$ or equivalently when r_{Aa} is below $2pq - 2pq\sqrt{\frac{\chi_1^2(\alpha)}{n}}$.

The acceptance region for HWE can thus be given in terms of the relative heterozygote frequency by:

$$(2pq - 2pq\sqrt{\frac{\chi_1^2(\alpha)}{n}} \leq r_{Aa} \leq 2pq + 2pq\sqrt{\frac{\chi_1^2(\alpha)}{n}}) \tag{2.4}$$

[3] 100(1-α) percentile of a χ_1^2 distribution

We now show the ternary plot of 1566 SNPs for the population we are studying (146 individuals).

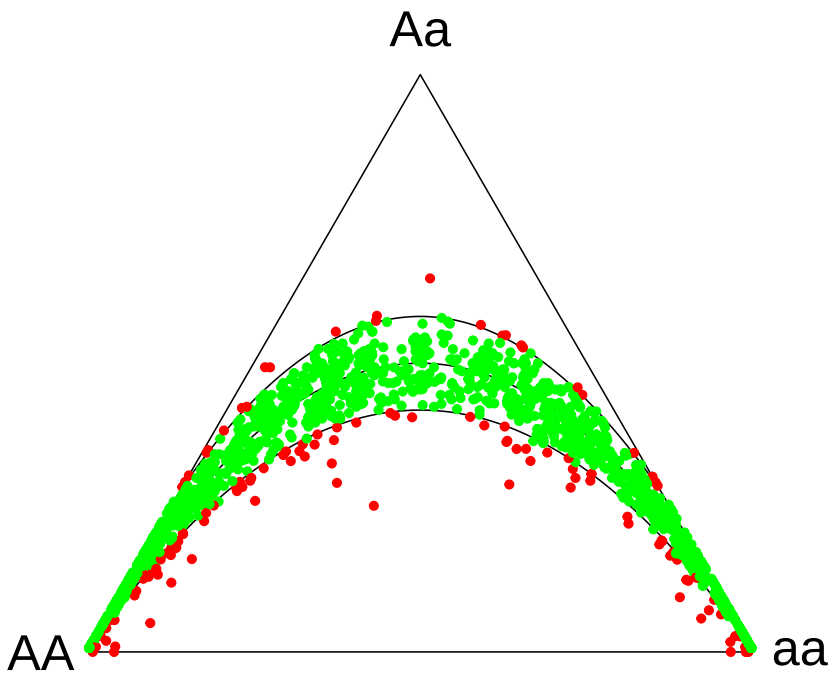

Figure 2.6: Ternary plot of the 1566 markers with the acceptance region corresponding to a Chi-square test

Figure 2.6 shows the ternary plot where markers in red are significant and markers in green are non-significant. They are colored automatically according to the result of a chi-square significance test.

In this case we find 138 significant markers that correspond to 8.81% of the total.
Note that HWE is refused in general, more frequently, because of heterozygote dearth rather than heterozygote excess.

HWE for cases and controls

Some investigators think that HWE is more likely to hold for controls than for cases. Cases may be subject to differential survival because of the disease and this could have distorted the HW proportions. For this reason, we are going to make an analysis of HWE stratified for cases and controls.

HWE for cases

We want to check if HWE holds in cases population (47 individuals).

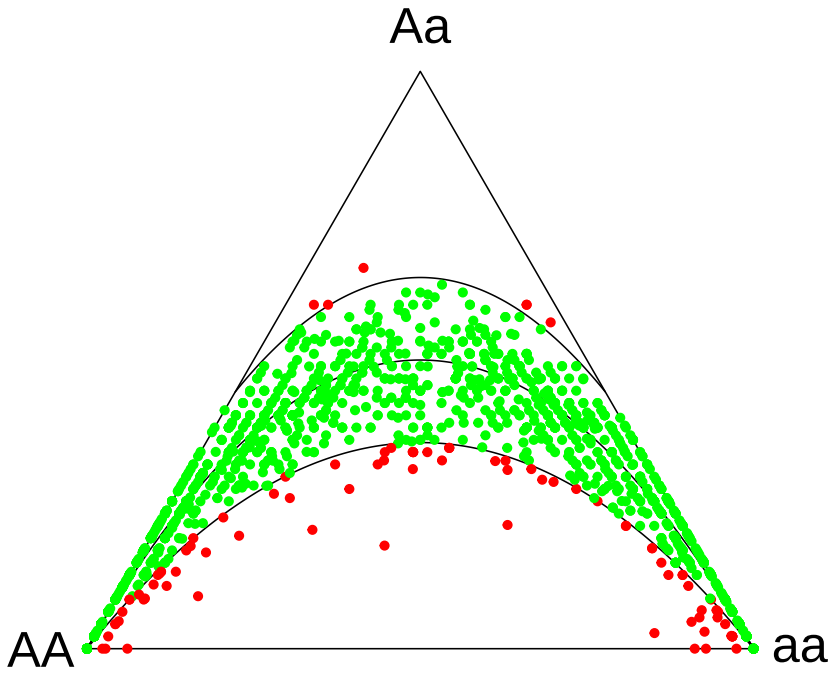

Figure 2.7: Ternary plot of the 1566 markers in Cases with the acceptance region corresponding to a Chi-square test

We find 92 significant markers, corresponding to 5.87% of the total.

HWE for controls

We check now if HWE holds in controls populations (99 individuals).

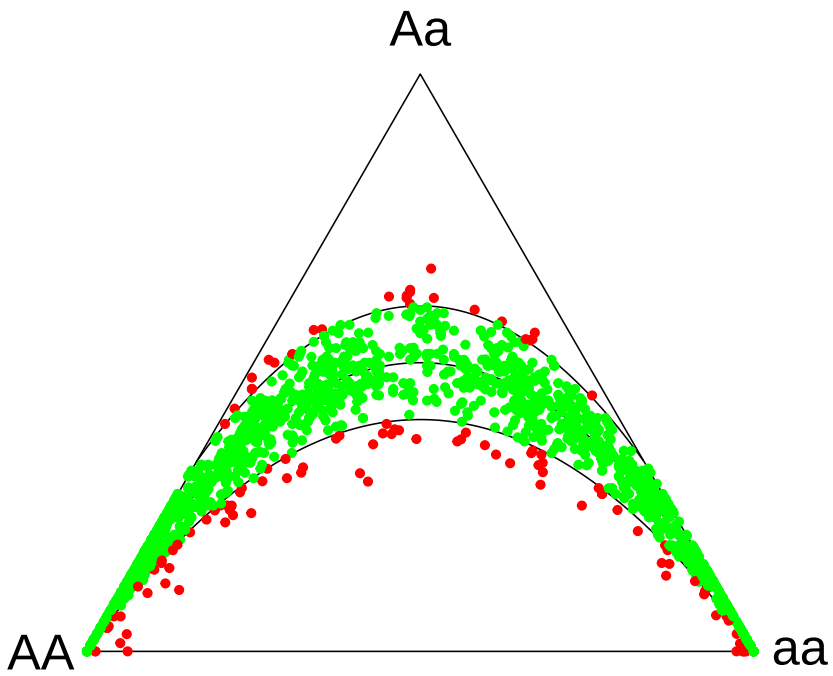

Figure 2.8: Ternary plot of the 1566 markers in Controls with the acceptance region corresponding to a Chi-square test

We find 113 significant markers, corresponding to 7.21% of the total.

We have not found evidence for more disequilibrium in cases but in controls. This result is reasonable because the sample size of controls doubles the sample size of cases and so there is more power to detect effects in controls.

	sample size	significant markers	% significant
Cases	47	92	5.87%
Controls	99	113	7.21%
Total	146	138	8.81%

Table 2.4: Results of testing for HWE

3 Genetic Association Analysis

As we mentioned, our goal is to investigate associations between markers and a trait (in this case: colon cancer's disease). The markers we are studying (1566) are bi-allelic polymorphisms. For each marker, we can test for association with the disease using different models.

3.1 General Theory

In this section we are going to present different statistical tests for marker-disease association. We subsequently describe four different statistical tests in the subsections below: the codominant test, Fisher's exact test, the alleles test and the Cochran Armitage trend test.

3.1.1 Codominant Test

We may consider the variable Y_i that takes value 1 if the individual i has the disease and value 0 if the individual does not have the disease. Further, as our markers are bi-allelic polymorphisms we can only have three different genotypes for each marker (AA, Aa, aa).

Individuals are organized into genotypic contingency tables according to their marker and disease status. Thus, we classify cases and controls according to their genotypes.

	AA	Aa	aa	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

Table 3.1: Genotypic contingency table

We want to test the null hypothesis of no effect of a marker on the trait.

We construct then a test of independence between the disease and genotype where the null and alternative hypothesis are the following:

$$H_0: \ P(Y = 1 \mid AA) = P(Y = 1 \mid Aa) = P(Y = 1 \mid aa)$$
$$H_1: \ \text{At least one pair of probabilities is different}$$

The test statistic is:

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{3.1}$$

where o_{ij} is the observed count and e_{ij} is the expected count cell (row total*column total / N).

Under H_0, we have $X^2 \sim \chi^2_2$. So, X^2 follows a chi- square distribution with two degrees of freedom (df) where df is given by $(nc-1)*(nr-1)$, with nc = number of columns and nr = number of rows of the contingency table.

3.1.2 Fisher's Exact Test

Fisher's exact test is a statistical significance test used in the analysis of contingency tables. It is named after its inventor, R. A. Fisher [1].

His motivating example was as follows. A British woman claimed to be able to distinguish whether milk or tea was added to the cup first and in order to test this claim she was given eight cups of tea. In four of them, tea was added first and in the other four, milk was added first. The cups were presented to her in a randomized order and she was told that there were four cups of each type.

For the contingency table, Fisher noted that under the null hypothesis of independence, if you assume fixed marginal frequencies for both the row and column marginals, then the hypergeometric distribution characterizes the distribution of the cell counts. This fact enables us to calculate an exact p-value rather than rely on an approximation that becomes exact in the limit as the sample size grows to infinity.

The p-value of the test is obtained by summing the probabilities of all possible tables that have the same margins as the observed table, but deviate as much or more from independence.

Ideally, we would use exact p-values all of the time. In practice, however, it is not possible because the algorithms in Exact Tests might break down as the size of the data set increases, as they require more computation time.

3.1.3 Allele test

This test aims to identify significant differences in allelic proportions between case and controls items. One supposes that alleles are binomially and independently sampled from cases and controls with probabilities p_{D_A} and p_{H_A} corresponding to the proportions of the A allele in the two sub-groups.

	A	a	Total
Cases	$D_A(=2r_0+r_1)$	$D_a(=2r_2+r_1)$	2r
Controls	$H_A(=2s_0+s_1)$	$H_a(=2s_2+s_1)$	2s
Total	$N_A(=2n_0+n_1)$	$N_a(=2n_2+n_1)$	2N

Table 3.2: Allelic contingency table

[1] Sir Ronald Aylmer Fisher(17 February 1890 - 29 July 1962) was an English statistician, evolutionary biologist, geneticist, and eugenicist

We want to test:

$$H_0 : \ p_{D_A} = p_{H_A}$$
$$H_1 : \ p_{D_A} \neq p_{H_A}$$

According to the null hypothesis, alleles are sampled from the same general population and allele frequency are the same for cases and controls.

Generally, we consider the following statistic

$$Z_A = \frac{\widehat{p}_{D_A} - \widehat{p}_{H_A}}{\sqrt{N\widehat{p}(1-\widehat{p})}} \tag{3.2}$$

where $\widehat{p}_{D_A} = \frac{D_A}{2r}$, $\widehat{p}_{H_A} = \frac{H_A}{2s}$ and $\widehat{p} = \frac{N_A}{2N}$

Under H_0, the p value can be calculated by considering that Z_A follows a Gaussian $N(0,1)$ distribution. When the sample size is small, a Fisher exact test leads to an exact computation of the p-value as we mentioned before.

We can also use the equivalent Pearson [2] statistic:

$$X^2 = (Z_A)^2 \tag{3.3}$$

Using this statistic, the p value is calculated by considering that $X^2 \sim \chi^2_1$.

3.1.4 Cochran-Armitage test

In case-control studies evaluating association between a candidate allele and a disease, allele and trend statistics are asymptotically equivalent when HWE holds. However, Sasieni [3] demonstrates that when HWE is not satisfied, the allele test is not valid and the Cochran-Armitage (CA) trend test can be used.

The trend test is based on the linear regression model.
If we define X as the number of A alleles and Y codes for cases and controls, i.e., Y = 1 for cases and Y = 0 for controls, then

$$P(Y = 1 \mid X) = \beta_0 + \beta_1 X + \varepsilon$$

We can use ordinary linear regression to estimate β_1 and test:

$$H_0 : \ \beta_1 = 0$$
$$H_1 : \ \beta_1 \neq 0$$

An alternative method is to treat X as a scaled (0,1,2) random variable giving the number of A alleles and compare the means of X in the two groups. By definition, sample means of X in the two groups and overall are

$$\bar{X}_{cases} = \frac{2r_0 + r_1}{r} = 2\hat{p}_{cases} \tag{3.4}$$

[2] Karl Pearson (27 March 1857 - 27 April 1936) was an influential English mathematician and biometrician

$$\bar{X}_{controls} = \frac{2s_0 + s_1}{s} = 2\hat{p}_{controls} \tag{3.5}$$

and

$$\bar{X} = 2\hat{p}. \tag{3.6}$$

Thus, testing $p_{cases} = p_{controls}$ is equivalent to testing that the means of X are equal in the two groups:

$$H_0 : E(X \mid case) = E(X \mid control)$$
$$H_1 : E(X \mid case) \neq E(X \mid control)$$

It can be shown that, under H_0,

$$var(\bar{X}_{cases} - \bar{X}_{controls}) = var(X)\left(\frac{1}{r} + \frac{1}{s}\right) \tag{3.7}$$

Further, var(X) can be estimated from the sample variance of the data by

$$\widehat{Var}(X) = \frac{4n_0 + n_1 - n\bar{X}^2}{rs} \tag{3.8}$$

As a result, we have that the test statistic for the trend test is:

$$Z = \frac{\bar{X}_{cases} - \bar{X}_{controls}}{\sqrt{\frac{4n_0+n_1-n\bar{X}^2}{rs}}} \tag{3.9}$$

Under H_0, $Z \sim N(0,1)$. Alternatively, $Z^2 \sim \chi^2_1$.

3.2 Example of each test for the first marker

In order to show clearly how the tests explained above work, we are going to do all the tests for the first SNP of the database.

We choose for all of them $\alpha = 0.05$ as the *significance level*.
Statistical significance is the low probability that an observed effect would have occurred due to chance. In any experiment or observation that involves drawing a sample from a population, there is always the possibility that an observed effect would have occurred as a result of just chance (or sampling error) alone. But if the probability of such chance effects is less than a pre-determined threshold (in this case 0.05), then an investigator can conclude that the observed effect actually reflects the characteristics of the population rather than just sampling error.

3.2.1 Codominant Test for the first SNP

We are going to do an example of the codominant test for the first marker.
First, we may be interested in knowing how many homozygous AA, heterozygous Aa and homozygous aa do we have for the first marker.

	# individuals
AA	72
Aa	60
aa	13

Table 3.3: Genotype counts for the first SNP

Next, we need to construct the genotypic contingency table for this marker. It is important to realize that there is one individual that does not have any genotype described for this marker but we already managed with missing values before.

	AA	Aa	aa
cases	27	16	3
controls	45	44	10

Table 3.4: Genotypic contingency table for the first SNP

The *chisq.test* function in R performs chi-squared contingency table tests and gives us the p-value, χ^2 statistic and the expected counts under the null hypothesis.

X-squared	df	p-value
2.2663	2	0.322

Table 3.5: Results of the codominant test for the first SNP

As the p-value > 0.05 (equivalently, the *X-squared statistic* < 5.99), we can't reject the null hypothesis of no effect of this marker to the disease.

A warning message is shown when performing this function in this case:
Warning: Chi-squared approximation may be incorrect.
This is because one of the expected counts is below 5 and therefore the chi-square approximation may not be right.
In these cases a good option would be to perform the exact test which will be done for the first SNP in the next subsection.

3.2.2 Fisher's exact test for the first SNP

To perform the Fisher's exact test we also need the genotypic contingency table 3.4 of the first marker.

The *fisher.test* function in R performs Fisher's exact test for testing the null of independence of rows and columns in a contingency table with fixed marginals.

p-value
0.3463

Table 3.6: P-value for the exact test for the first SNP

As the p-value > 0.05, we can't reject the null hypothesis of no effect of this marker to the disease. Note that the χ^2 test and the exact test almost have the same p-value and thus lead to the same conclusion.

3.2.3 Allele test for the first SNP

What we need first, is to construct the allelic contingency table for the first marker.

	A	a
Cases	70	22
Controls	134	64

Table 3.7: Allelic contingency table for the first SNP

As we mentioned while explaining the allele test theory, we can use again the Pearson statistic X^2 that follows a chi-square distribution with one degree of freedom χ^2_1, which can be calculated again with the *chisq.test* function.

X-squared	df	p-value
2.1298	1	0.1445

Table 3.8: Results of the allele test for the first SNP

Again, as the p-value > 0.05, we can't reject the null hypothesis of no effect of this marker to the disease.

3.2.4 Cochran-Armitage trend test for the first SNP

In this subsection we are going to perform the Cochran-Armitage trend test for the first SNP. In this case we have not found any function in R that gives us the p-value or the statistic for this test. Thus, we have created a function that calculates the statistic for the trend test. Once we have the statistic we just have to compare it with the critical value $(N(0,1)(\alpha))$, for $\alpha = 0.05$

statistic
1.453423

Table 3.9: Statistic of the trend test for the first SNP

As the critical value $(N(0,1)(\alpha))$ with $\alpha = 0.05$ is 1.64 and the statistic has a lower value, we can't reject the null hypothesis,

3.3 Tests for all the markers

We have shown how to perform the different tests for one SNP. The next step is to do the same but for all the markers in order to find which markers are significant in each of the tests and so detect possible associations between these markers and the disease.

We thus describe the results of testing all the markers with the different tests we already introduced in previous sections.

We are going to present first the Bonferroni correction.

Bonferroni correction

The correction is based on the idea that if we have n independent significance tests at the α level and we want to guarantee that the overall significance test is still at this level, then we need to test each individual hypothesis at a statistical significance level of $\frac{1}{n}$ times what it would be if only one hypothesis were tested. So, if it is desired that the significance level for the whole family of tests should be α, then the Bonferroni correction would be to test each of the individual test at a significance level of $\frac{\alpha}{n}$.

Thus, if we use the Bonferroni correction the new significance level for each individual test is:

$$\alpha' = \frac{\alpha}{n} = \left(\frac{0.05}{1566}\right) \tag{3.10}$$

3.3.1 Codominant test for all the SNPs

In this subsection we will do the codominant test for all the markers of our data frame.

We already create a function that performs this test for one SNP and we used it for the first marker. Using this function we get the results of the codominant test for all the markers. It gives us the p-value of this test for each of the markers so we can compare them with the significance level which is 0.05 and also with the significance level defined by the Bonferroni correction, which is $\frac{0.05}{1566}$.

In order to emphasize the significant markers we apply -log() function to the p-values. Red and green lines in the plot express the significance level without taking into account the Bonferrioni correction and using the Bonferroni correction respectively.

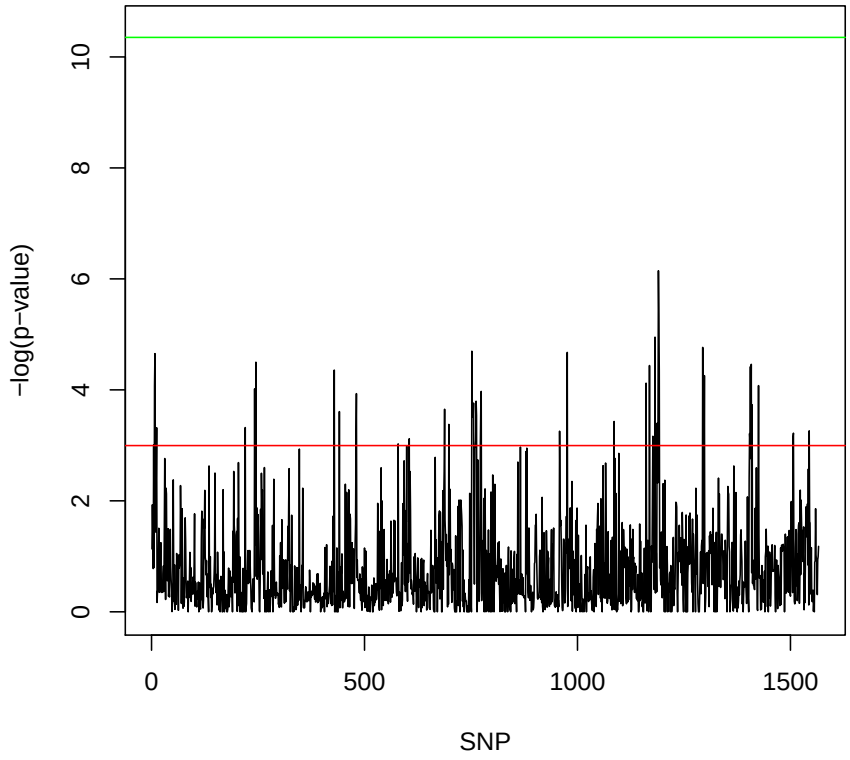

Figure 3.1: Plot of -log(p-value) of the chi-square test for each SNP of the database

Figure 3.1 shows the plot of -log(p-value) of the codominant test for each SNP of the database. Further, the plot contains a horizontal line that allows us to see whether a marker is significant for this test or not. If the p-value of a marker is lower than the significance level which is 0.05 (equivalently, -log(p-value) is higher than -log(0.05)) then we can reject the null hypothesis of no effect of this marker to the disease and that means some associations could exist between this marker and the disease.

After doing the Codominant test for all the markers, we can conclude that there are 49 significant markers for this test that represent a 3.12% of the total. Also notice that the most significant marker is the 1190 and that using the Bonferroni correction, we do not find any significant markers.

If the null hypothesis is true for all genetic markers then we may expect a 5% significant tests by chance. We find nearly 3% which is below 5% so apparently data marker information is not related to the disease.

Even so, we observe clustering of significant markers in Figure 3.1. We check for randomness of the sequence of significant and non-significant results by using the function *runs.test* of *tseries*-package [4].

statistic	p-value
-12.121	2.2e-16

Table 3.10: Results of the runs test for randomness

Table 3.10 gives us the statistic and the p-value of the runs test for the sequence of significant and non-significant results for the codominant test.
As p-value (= 2.2e-16) < 0.05, we can reject the null hypothesis of randomness of the sequence.

Furthermore, it is important to know which markers are significant for the codominant test in order to compare results between the other tests that will be performed. These markers are the following 49.

```
5    7    8   12  220  242  245  429  441  480  481  579  605  688  698
752  753  754  755  756  761  762  763  772  773  774  958  976 1086 1161
1169 1177 1182 1187 1189 1190 1191 1192 1294 1298 1404 1405 1406 1408 1410
1425 1506 1507 1544
```

We can also evaluate the chi-square statistic distribution with a Q-Q plot. In statistics, a Q-Q plot is a probability plot, which is a graphical method for observing how close the probability distribution of a sample is to a theoretical distribution by plotting their quantiles against each other. In this case we want to see how close the distribution of the chi-square statistic is to the Chi-square distribution.

Take in mind that some markers only have two genotypes instead of three. In this cases the chi-square statistic follows a chi-square distribution with one degree of freedom instead of 2 ($X^2 \sim \chi^2_1$). For this reason, the Q-Q plot shown in the next Figure evaluates the chi-square statistic distribution where the chi-square statistic have been calculated for those markers that have the 3 genotypes. Thus, we compare this distribution with the chi-square with two degrees of freedom distribution. We have 1233 markers with 3 genotypes and 333 markers with 2 genotypes (the markers with only one genotype were removed from the database).

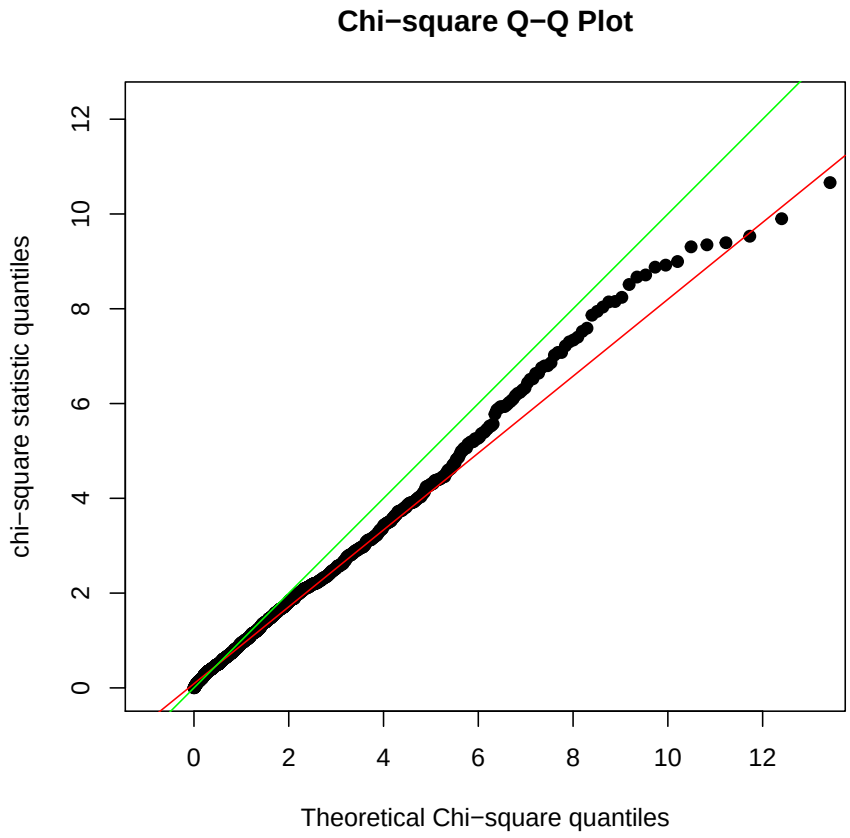

Figure 3.2: Chi-square Q-Q plot

As it is shown in Figure 3.2 the chi-square with two degrees of freedom distribution seems quite appropriate for our data. Even so, at the upper tail of the distribution, observed values are larger than expected for a χ^2_2. These are, indeed, the significant tests (chi-square statistic from 5.99).

3.3.2 Fisher's exact test for all the SNPs

Again, we have already created a function that performs this test for one SNP so we can use it to do the test for all of them.

We now give graphically the p-value of the Fisher's exact test for each of the markers and we compare it with the significance level. In order to emphasize the significant markers we apply -log() transformation to the p-value.

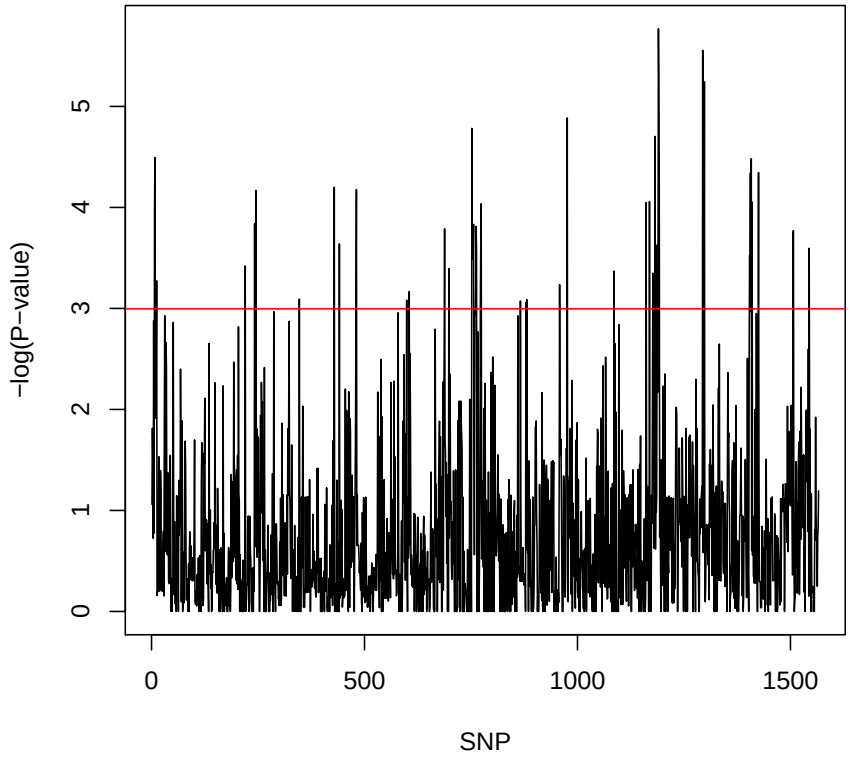

Figure 3.3: Plot of -log(p-value) of the exact test for each SNP of the database

Figure 3.3 shows the plot of -log(p-value) of the Fisher's exact test for each of the 1566 SNPs. Further, the plot contains an horizontal line that allows us to see whether a marker is significant for this test or not. If the p-value of a marker is lower than the significance level (equivalently: if -log(p-value) is higher than -log(0.05), then we can reject the null hypothesis of no effect of this marker to the disease and that means some associations could exist between this marker and the disease.

We can conclude, this time, that there are 52 significant markers for this test that represent a 3.32% of the total. Notice that in this case the most significant marker is also the 1190.

We find in this case 3.32% of significant markers which is below 5% so apparently data marker information is not related to the disease.

Again, we observe clustering in Figure 3.3. Moreover, if we compare Figures 3.3 and 3.1 we see clustering in the same regions.

We are going to perform again the runs test in order to discard randomness of the sequence of significant and non-significant results given by the exact test.

statistic	p-value
-11.2879	2.2e-16

Table 3.11: Results of the runs test for randomness

Table 3.11 gives us the statistic and the p-value of the runs test for the sequence of significant and non-significant results for the exact test.
As p-value (= 2.2e-16) < 0.05, we can reject the null hypothesis of randomness of the sequence.

In addition, it is interesting to realize that there are 47 markers that are significant for both tests. These markers are the following:

```
7    8   12  220  242  245  429  441  480  481  605  688  698  752  753
754  755  756  761  762  763  772  773  774  958  976 1086 1161 1169 1177
1182 1187 1189 1190 1191 1192 1294 1298 1404 1405 1406 1408 1410 1425 1506
1507 1544
```

3.3.3 Allele test for all the SNPs

Using the function we created to perform the allele test for one SNP, we calculate the p-value for all the markers of the database.

In order to emphasize the significant markers, we apply -log() function to the p-value as we did in previous subsections and we show the results in the following plot.

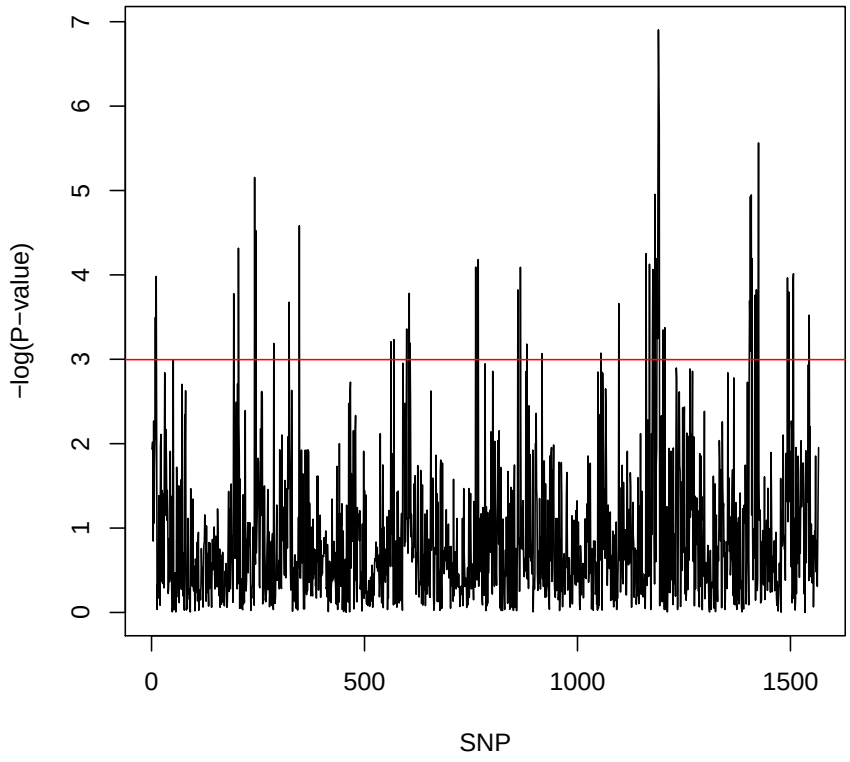

Figure 3.4: Plot of -log(p-value) of the allele test for each SNP of the database

Figure 3.4 shows the plot of -log(p-value) for the allele test for each of the 1566 SNPs. Further, the plot contains a horizontal line that allows us to see whether a marker is significant for this test or not.

We have 49 significant markers for the allele test that represent a 3.12% of the total. Notice that the most significant marker is again the 1190.

Let's test for randomness of the sequence of significant and non-significant results given by the allele test.

statistic	p-value
-7.9318	2.16e-15

Table 3.12: Results of the runs test for randomness

Table 3.12 gives us the statistic and the p-value of the runs test for the sequence of significant and non-significant results for the allele test.
As p-value (= 2.16e-15) < 0.05, we can reject the null hypothesis of randomness of the sequence.

From the 47 markers that we already separated because of being significant for the previous tests, only 19 are also significant for the allele test. These markers are the following:

```
242  245  605  761 1161 1169 1177 1182 1187 1189 1190 1191 1192 1404 1405
1406 1408 1410 1425
```

3.3.4 Cochran-Armitage trend test for all the SNPs

Using the function we already create to perform the trend test for one SNP, we calculate the statistic of the trend test for each SNP of the database.

The next figure contains the statistic value for all the SNPs and also two red lines that mark the two critical values $(N(0,1)(0.05))$ and $(N(0,1)(0.95))$ because we are considering a two-sided test.

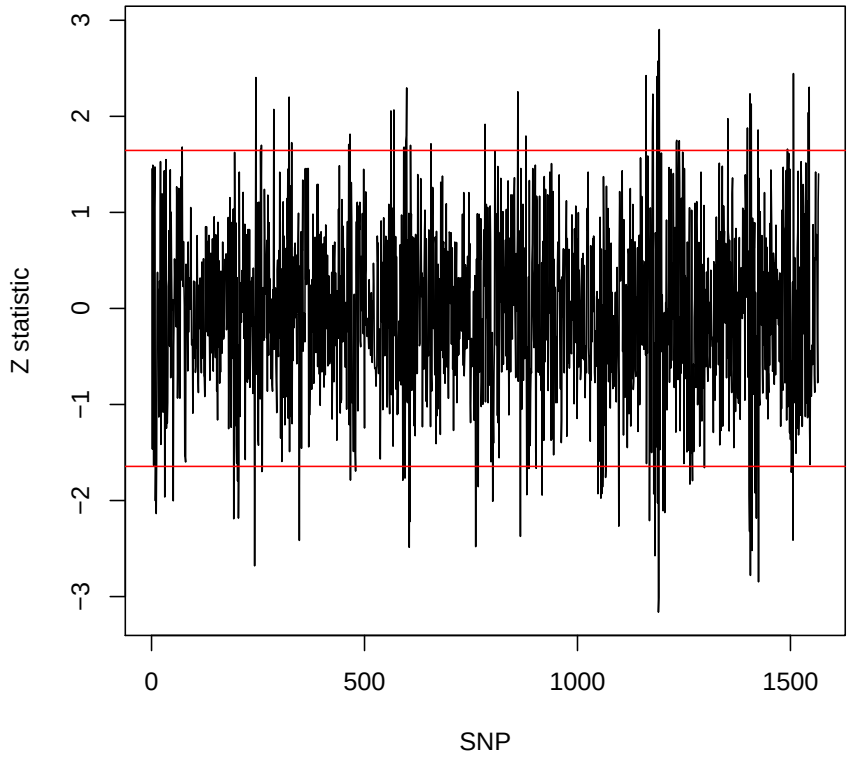

Figure 3.5: Plot of the Z statistic of the Cochran-Armitage trend test for each SNP of the database

Figure 3.5 shows the statistic value for each SNP.
Performing this test we find 91 significant markers which is a number much larger than the ones found for the other tests. The significant markers represent a 5.81% of the total which is more than what we would expect due to chance.

We are going to test for randomness of the sequence of significant and non-significant results of the trend test.

statistic	p-value
1.8888	0.05892

Table 3.13: Results of the runs test for randomness

As Table 3.13 shows, the p-value $(=0.059) > 0.05$. In this case we can not reject the null hypothesis of randomness of the sequence of results.

How many markers are significant for all the tests?

Only 19 markers are significant in all the tests. These markers are the same ones that were significant in the codominant test, the exact test and the allele test.
These 19 markers are the following:

```
242  245  605  761 1161 1169 1177 1182 1187 1189 1190 1191 1192 1404 1405
1406 1408 1410 1425
```

3.4 Results

In this section we characterize the markers that are significant in all tests in terms of their position, minor allele frequency, HWE status and degree of significance in the codominant test. Results are shown in Table 3.14.

Marker	MAF	p-value HWE	p-value codominant test
242	0.34	0.53	0.02
245	0.35	0.07	0.01
605	0.47	0.18	0.04
761	0.42	0.50	0.05
1161	0.13	0.57	0.02
1169	0.11	0.02	0.01
1177	0.17	0.84	0.04
1182	0.47	0.11	0.01
1187	0.18	0.82	0.03
1189	0.28	0.29	0.03
1190	0.33	0.27	0.00
1191	0.25	0.40	0.00
1192	0.26	0.92	0.01
1404	0.06	0.84	0.04
1405	0.08	0.66	0.05
1406	0.06	0.76	0.01
1408	0.08	0.00	0.01
1410	0.07	0.76	0.02
1425	0.13	0.79	0.02

Table 3.14: Position, MAF, HWE status and degree of significance for the codominant test for the 19 markers that are significant in all the tests

Table 3.14 shows that the markers that are associated to colon cancer tend to occur in blocks of more or less consecutive markers. At least two blocks of colon cancer related markers can be distinguished. These are the markers with position 1161 through 1192 and 1404 through 1410. The first block contains 9 markers and the second one 4 markers. 4 markers in the first block and three markers in the second block are consecutive. The second block consists of markers with a low MAF, in the range of 0.06-0.08.

Further, we construct a table that contains the counts at each combination of significant and non-significant markers in the different tests. We do this in order to find out whether the tests are consistent between each other. E.g. there are 1512 markers that are not significant in both allele

test and codominant test, 5 markers that are significant in allele test but not in codominant test, 2 markers that are not significant in allele test but in codominant test and finally, 47 markers that are significant in both tests.

Counts		Codominant test		Exact test		Allele test		Cochran Armitage	
		Not Sig	Sig	Not Sig	Sig	Not Sig	Sig	Not Sig	Sig
Codominant test	Not Sig	1517	0	1512	5	1490	27	1449	68
	Sig	0	49	2	47	27	22	26	23
Exact test	Not Sig 3	1512	2	1514	0	1491	23	1451	63
	Sig	5	47	0	52	26	26	24	28
Allele test	Not Sig	1490	27	1491	26	1517	0	1474	43
	Sig	27	22	23	26	0	49	1	48
Cochran-Armitage	Not Sig	1449	26	1451	24	1474	1	1475	0
	Sig	68	23	63	28	43	48	0	91

Table 3.15: Contingency table of the counts at each combination of significant and non-significant markers in the different tests

Table 3.15 shows that exact and codominant tests are very similar. That means they reach to the same conclusion for almost all the markers. Moreover, allele test and trend test also seem to be similar but they differ more with the other two tests.

In addition, we show in the next table the results of the runs test for the sequencies of significant and non-significant results of each test.

	statistic	p-value
Sequence of results codominant test	-12.121	2.2e-16
Sequence of results exact test	-11.2879	2.2e-16
Sequence of results allele test	-7.9328	2.16e-15
Sequence of results Cochran-Armitage test	1.8888	0.05892

Table 3.16: Results of the runs test

We can see in Table 3.16 that the sequence of significant and non-significant markers of the codominant test, the fisher test and the allele test are not considered random. The randomness test for the the Cochran-Armitage test is not significant, but it's p-value is so close to the 5% threshold that some evidence against a random distribution of significant markers is present.

4
Simulations

In the previous chapter we found two distinguished clusters in the results of the tests. The first cluster consists of markers with position 1161 through 1192 and the second one consists of markers with position 1404 through 1410. These markers are significant in all tests and furthermore, we found that the results of the codominant test, the exact test and the allele test were not random.

In this chapter we are going to repeat some of the tests with simulated data marker. The objective of this chapter is to deffend the idea that the clusters of significant markers found in all the tests for the real marker data that we are studying are not random, but related in some way to the colon cancer disease.

How do we simulate data marker?

For each of the 1566 SNPs, we calculate their genotype frequencies which gives us the proportion of the three genotypes in our population (146 individuals). Once we have this probabilities, we construct, for each SNP, a sample of size 146 where the elements(genotypes) will appear with the given probabilities. As we do this for all the SNPs, what we get is a new database of 1566 markers and their genotypes for the 146 individuals. The simulated database matches the original database in the sense that the markers have the same genotype and allele frequencies, but are unrelated to the disease indicator.

4.1 Codominant test for the simulated data

In this section we are going to repeat the codominant test using the simulated data we have created.

We give graphically -log(pvalue) in order to emphasize the significant markers. The following figure shows the significant markers that we get whith the genotypes for each SNP given randomly for all the individuals.

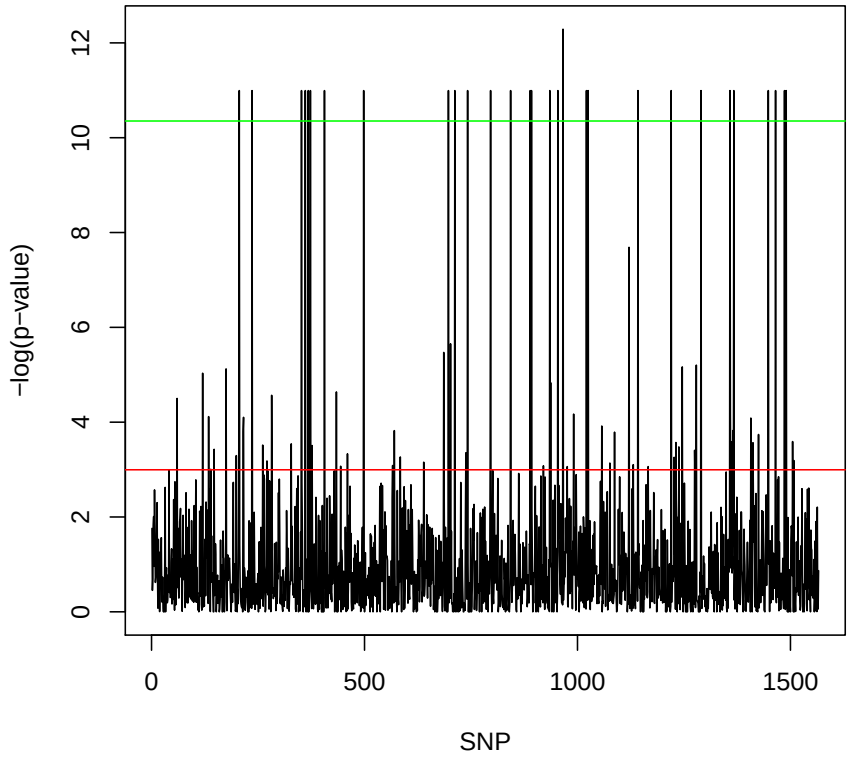

Figure 4.1: Plot of -log(p-value) of the chi-square test for each SNP of the simulated database

After doing the Codominant test for the simulated data marker, we can conclude that there are 80 significant markers that represent a 5.1% of the total. We find more significant markers in the simulated data than in our database. In fact, the percentage of significant markers is very close to what we would expect by chance (5%).

We do not see clustering in Figure 4.1 so the clustering we found seems to be characteristic of our marker data.

Although it doesn't seem to present clustering, we are going to perform the runs test and see what it says about the randomnes of the sequence of significant and non-significant markers.

statistic	p-value
0.0454	0.9638

Table 4.1: Results of the runs test for randomness

As we see in the table, the p-value (=0.96) > 0.05 so we can not reject de null hypothesis of randomness of the sequence.

As a conclusion we say that if we generate random genotypes with same genotype frequencies that we have in our data marker, we do not see clustering in the results and furthermore, the sequence

of significant and non-significant markers is considered random.

4.2 Cochran-Armitage test for the simulated data

In this section we will perform the allele test for each SNP using the simulated data.

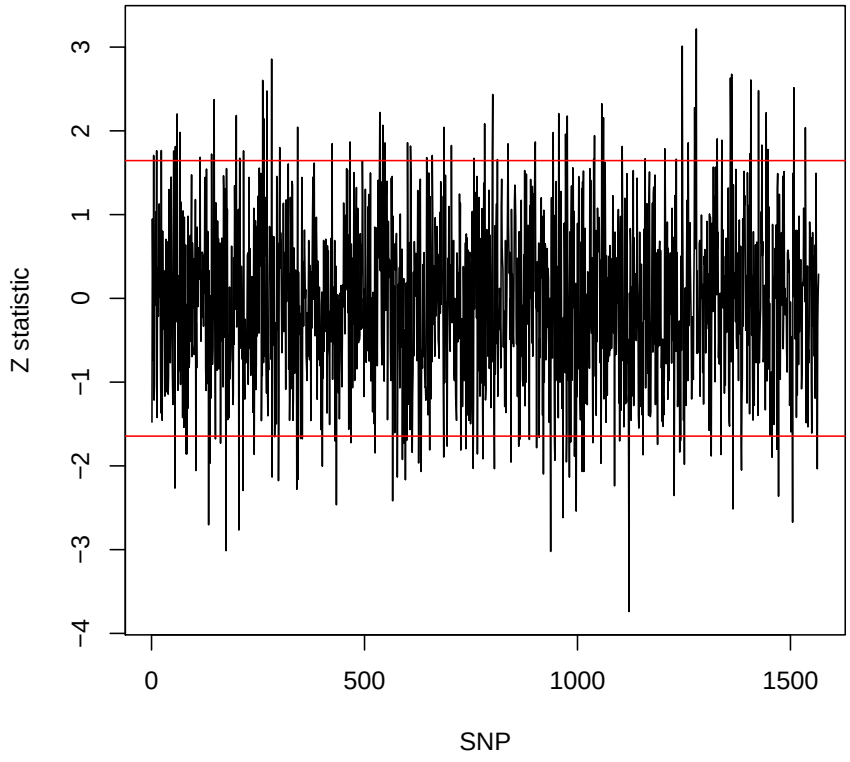

Figure 4.2: Plot of -log(p-value) of the Cochran-Armitage test for each SNP of the simulated database

After doing the Cochran-Armitage test for the simulated data marker, we can conclude that there are 148 significant markers that represent a 9.45% of the total. We find more significant markers in the simulated data than in our database. In fact, we find more markers than we would expect by chance.

We do not see clustering in Figure 4.2 so the clustering we found was characteristic of our data marker.

Although it doesn't seem to present clustering, we are going to perform the runs test and see what it says about the randomnes of the sequence of significant and non-significant markers.

statistic	p-value
1.1792	0.2383

Table 4.2: Results of the runs test for randomness

As we see in the table, the p-value (=0.2383) > 0.05 so we can not reject de null hypothesis of randomness of the sequence.

The two test we have performed with the simulated data marker do not present clustering in the two regions we previously identified. Instead, we can say that the sequence of significant and non-significant markers is random in this case.

5 Conclusions and discussion

In this project we have analized a database containing marker information of 146 individuals, 47 of them suffering from colon cancer. The goal of this project was determine possible associations between markers and the disease trait.

We initially had in our database the genotypes of these individuals for 1685 markers (SNPs). We decided to remove 119 markers from the database either because they presented too much missings or because they were monomorphic markers.

We first calculated the minor allele frequency (MAF) for each of the 1566 markers that remained in our database after the removal. Calulating the MAF helped us to see the degree of variance these markers presented. We found that markers with MAF in the range of 0.00 - 0.01 were relatively more common, so a lot of markers were not very informative.

We also checked for Hardy Weinberg Equilibrium in order to discard genotyping error. Using the $HWTernaryPlot$ function we only found 138 significant markers which represent a 8.81% of the total. Further, we tried to investigate if there were more significant markers in cases population than in controls population. We did this because some investigators think that Hardy Weinberg Equilibrium is more likely to hold in cases than in controls. We actually didn't find evidences of more disequilibrium in cases than in controls but that is probably because the sample size of controls doubles the sample sizes of cases.

After doing an exhaustive description of our dataset, we started to test for associations between the markers and the disease using four different models: The codominant test, the exact test, the allele test and the Cochran-Armitage test.
These four tests do not yield identical results. We found strong agreement between the test results of the codominant test and the exact test, and also between the alleles test and the CA test. We found relatively more discrepancies between these two sets (see 3.15).

The results of this tests are shown in the following table.

Test	# significant markers	%significant	p-value runs test
Codominant test	49	3.12%	2.2e-16
Exact test	52	3.32%	2.2e-16
Allele test	49	3.12%	2.16e-15
Cochran-Armitage test	91	5.81%	0.059

Table 5.1: Results for each of the tests

As we can see in Table 5.1, the percentage of significant markers is somewhat lower than the percentage expected by chance alone (5%). Only the Cochran-Armitage test gave a slightly higher percentage of significant markers.

Although we haven't found more significant markers than expected due to chance we have found clustering in significant markers in the results of all the tests. Furthermore, the clustering has been found in the same regions.
We have performed a test of randomnes to see if the sequences of significant and non-significant markers can be considered random or not. We can see the results of this test in Table 5.1. For three of the tests the sequence is not considered random so it is a signal that the clustering does possibly exist. The randomness test for the the Cochran-Armitage test is not significant, but it's p-value is so close to the 5% threshold that some evidence against a random distribution of significant markers is present.

We also have determined which markers are significant for all the tests. We have found 19 markers. These markers are the following:

```
242  245  605  761 1161 1169 1177 1182 1187 1189 1190 1191 1192 1404 1405
1406 1408 1410 1425
```

Two distinguished blocks of colon cancer related markers can be determined in this sequence of 19 markers. These are the markers with position 1161 through 1192 and 1404 through 1410. The second block can be considered less informative because it contains markers with a low MAF (in the range of 0.06 - 0.08).

Finally, by simulating new data marker we have seen that the two groups of significant markers that have been found are not found when the data marker is randomly determined. Furthermore, the sequences of significant and non-significant markers found after performing the codominant test and the Cochran-Armitage test in this case can be considered random. Thus, we conclude that some associations apparently exist between these two groups of markers and colon cancer disease.

Because we have done many statistical tests, we cannot exclude that the detected clusters represent a chance effect. It would be interesting to see if our list of 19 significant markers again turns up significant for new sample of cases and controls.

The observed clustering between significant markers may be due to correlation between neighbouring markers (linkage disequilibrium). This could be studied in future work.

Bibliography

[1] HardyWeinberg: Graphical tests for Hardy-Weinberg equilibrium, Jan Graffelman, 2013, R package version 1.5.2, http://CRAN.R-project.org/package=HardyWeinberg.

[2] Graffelman, J. and Morales-Camarena, J. *Graphical tests for Hardy-Weinberg Equilibrium based on the ternary plot.* Human Heredity 2008; 65(2): 77-84.

[3] Sasieni PD: *From genotypes to genes: Doubling the sample size.* Biometrics 1997; 53: 1253-1261.

[4] tseries: Time Series Analysis and Computational Finance, Adrian Trapletti and Kurt Hornik, 2013, R package version 0.10-32., http://CRAN.R-project.org/package=tseries.

[5] Boris Freidlin, Gang Zheng, Zhaohai Li and Joseph L. Gastwirth. *Trend Tests for Case-Control Studies of Genetic Markers: Power, Sample Size and Robustness.* Human Heredity 2002; 53: 146-152.

[6] Nan M. Laird and Christoph Lange: *Statistics for Biology and Health. The Fundamentals of Modern Statistical Genetics.* Springer 2011.

[7] Peter Dalgaard: *Statisfics and Computing. Introductory Statistics with R.* Springer 2008.