

A Strategy for Detecting Multiple Trait Loci in Disease Association Studies

Caroline Uhler and Anna-Sapfo Malaspinas

May 2008

Abstract

Rapid research progress in genotyping techniques have allowed large genome-wide disease association studies. Existing methods often focus on determining associations between single loci and the disease. However, most diseases involve complex relationships between multiple loci and the environment. Here we describe a method for finding interacting loci by combining the traditionally used single-locus search with a search for multiway interactions on contingency tables. First, we develop an extended Fisher's exact test for multidimensional contingency tables. To do so, we introduce toric ideals and construct a Markov chain on the space of multidimensional contingency tables with fixed margins. Second, we test our methods on simulated data, showing that we can detect interacting loci where single locus methods fail to do so.

1 Introduction

In this section we will give an introduction to relevant biological concepts and mathematical tools, on which we will build our model for detecting interactions between loci. We will first explain the problem studied in this paper from a biological point of view. Then, we will discuss the term 'interaction' and finally review two widely used tests for detecting interactions in contingency tables, namely the chi-squared goodness-of-fit test and Fisher's exact test.

1.1 General background

Two individuals taken at random have almost identical DNA sequences. Indeed, on average only about one in every 1,200 bases differ between two individuals (The International HapMap Consortium (2003)), which represents a total of about 600,000 nucleotides over the whole genome. This small amount of variation is thought to account for most of the trait variation among humans, including diseases. The most common type of genetic variation are differences in individual base pairs. A site that

differs between individuals in the human genome is called *single-nucleotide polymorphism* (SNP). Among the four possible nucleotides of the DNA (A, C, G, T) typically only two nucleotides commonly occur at a particular site and can therefore be encoded by 0 and 1, respectively. The sequence of SNPs on a chromosome is called *haplotype*. However, each individual possesses two copies of each chromosome (aside from the sexual chromosomes) and therefore two haplotypes, which might be unequal. Collecting haplotype data empirically is prohibitively expensive. Therefore, often only the genotype data is collected instead. The *genotype* is a superposition of the two haplotypes, where the binary characters of the two corresponding SNPs are added together. So sites with values 0 and 2 are homozygous and denote the combination 0-0 and 1-1, respectively. A site with value 1 is heterozygous and denotes either the combination 0-1 or 1-0.

Many diseases have a genetic component. Some are monogenic and are associated with variation in a single gene. For example, in the well-known case of cystic fibrosis, a single mutation in the CFTR gene, when inherited from both parents, is sufficient to make an individual diseased. The CFTR gene encodes a ion channel whose functionality is altered by the presence of a single mutation in the gene. An absence of at least one functional chloride channel leads to progressive disability due to multisystem failure.

Genes related to monogenic diseases are often easy to map on the genome. However, most common diseases are multifactorial. The factors can be of both genetic and environmental origin. More than one gene can be causative for a disease and only the combination of particular variants will make an individual diseased. Each genetic variant independently can have a very low effect, undetectable by a standard single locus approach. Multidomain proteins, alternative biochemical pathways, and other types of epistasis are possible modes of interaction of SNPs in order to produce a particular phenotype. Although we will not consider them directly in this study, environmental factors play an important role as well (e.g. diet, smoking, etc.). The most prevalent of these disorders include cancer, diabetes and hypertension. In other words, methods to detect SNPs associated with a disease are of extreme importance in order to gain a better understanding of the molecular basis of widely spread diseases with high morbidity.

Genome variation is hence expected to play an important role in disease association studies. The idea of such studies is to find genetic factors that are correlated with a disease by comparing the SNPs of non-diseased individuals to the corresponding SNPs of disease carriers.

Genome-wide studies have now become feasible by the combination of high-throughput genotyping methods, DNA chips holding 500'000 SNPs and large collections of well-phenotyped human samples. In other words, large data sets for population-based disease association studies are now available. It is therefore essential to develop computational methods to analyze this data productively.

Although great progress has been achieved in the last few years using simple linear models, those methods remain inefficient for large-scale data to detect interactions (Albrechtsen et al. (2007)). Recent studies have revealed the importance of interactions between multiple loci for various diseases (e.g. Marchini et al. (2005)). Various models of interaction have been presented in the past, as for example additive models or multiplicative models. The former model, assumes that the SNPs act independently, and a single marker approach seems to perform well. In multiplicative models SNPs interact in the sense that the presence of two (or more) variants have a stronger effect than the sum of the effects of each single marker. We will discuss similar models in Section 3), trying to cast them in a biological framework.

In this study, we will first reduce the possibly interacting SNPs to a small number by filtering the genome-wide SNPs with a single locus approach, following what has been suggested by Marchini et al. (2005). Indeed, they discuss a two-stage approach for performing multi-locus searches. First, all loci achieving some threshold in a single-locus search are identified. These loci do not necessarily need to give significant p-values under the single-locus search. These loci are then further examined for interactions between a given number k of loci by performing the χ^2 -test for all k -subsets of loci. For some models of interaction, Marchini et al. (2005) show that the two-stage approach outperforms the single-locus search and performs at least as good as when computing the χ^2 -score of all k -way interactions of all SNPs. This result suggests that a two-stage approach is reasonable in our case.

Single locus methods consider each SNP individually and test for association based on differences in allele frequencies between case and control individuals. A widely used method for a single-locus search is based on the χ^2 goodness-of-fit test and is described in the next subsection. Bonferroni corrections for the p-value are in general used to account for the large number of tests performed.

However, it is desirable to test various associations between a selection of markers by an exact test and without having to do computationally intensive permutation tests for all k -subsets. In Section 2 we will present how various associations within a selection of markers can be tested by extending Fisher's exact test on 2-dimensional tables to a test for association within multi-dimensional tables. This leads to exact tests for associations within multi-dimensional tables and does not require performing a permutation test for every k -subset of the data.

In what follows, we first discuss the term 'interaction' and then present two widely used tests for detecting interaction within 2-dimensional contingency tables. The χ^2 goodness-of-fit test is a standard method for testing for independence between classes in contingency tables. While the χ^2 goodness-of-fit test is an approximate test and therefore problematic with small counts in the table, an exact test is presented in Section 1.4.

1.2 Interaction between markers

It is important to note that the term ‘interaction’ has different meanings in different contexts. In a biological context, interaction between markers (or SNPs) is in general used as synonym for *epistasis*. Cordell (2002) gives a broad definition: “Epistasis refers to departure from ‘independance’ of the effects of different genetic loci in the way they combine to cause disease”. Epistasis is for example the result of a multiplicative effect between two markers in order to cause the disease.

In contrast, in a mathematical context interaction is used as synonym for *correlation*. Two markers are said to be interacting if they are correlated, i.e.

$$\mathbb{P}(\text{marker 1} = i, \text{marker 2} = j) \neq \mathbb{P}(\text{marker 1} = i)\mathbb{P}(\text{marker 2} = j).$$

In general, in association studies the goal is to find a set of markers that are correlated with the disease. However, the markers can be correlated with each other as well. Our hypothesis is that detecting this correlation might help understanding the type of interaction between the markers and might also result in a gain of power to detect the causative SNPs themselves.

Correlation between the markers can have different causes. One possibility is *epistasis*. This can be best understood in the extreme example of lethal combinations. Imagine that having marker 1 and marker 2 both in state 1 is lethal for an individual, but having just one of these two markers in state 1 is not lethal. In this case, the joint probability of both markers being in state 1 is 0, whereas the product of the two single events might be small but non-zero.

Another possible cause of correlation is *linkage disequilibrium*. The probability of having recombination between two SNPs that lie near to each other on a chromosome is small. So these SNPs are linked and therefore correlated. This correlation decreases with the distance between two SNPs and is measured by the linkage disequilibrium.

Finally, *non-random sampling* is another source of correlation between two markers. Usually, in disease association studies half of the individuals are diseased and half of them are not. This is crucial in order to have enough power to detect the SNPs associated with the disease. However, this procedure also induces correlation into the data as explained by the following example. Assume that only individuals with marker 1 and 2 being jointly in state 1 are diseased. So by the design of the experiment, the probability of having both markers jointly in state 1, (which is equal to the probability of being diseased) is 0.5. But the product of the two probabilities might be smaller. This example shows that non-random sampling can introduce high correlation, which might pose problems.

From now on we will use the term interaction as synonym for correlation. Correlation is usually measured by the χ^2 -statistic, as described in the following subsection.

1.3 χ^2 goodness-of-fit test

Assuming haplotype data, each marker (SNP) can be in $s = 2$ states and so the association between each single SNP and the disease can be summarized in a 2×2 contingency table (see Section 1.4). The number of observations in a particular class is denoted by a_{ij} . The test statistic T is given by

$$T = \sum_{i,j} \frac{(a_{ij} - E_{ij})^2}{E_{ij}},$$

where E_{ij} represents the expected number of observations in class (i, j) under the assumption that the null hypothesis is true. In our case, the null hypothesis consists of no association between a marker and the disease. So under H_0 we get:

$$E_{ij} = \frac{\sum_m a_{mj} \sum_n a_{in}}{\sum_{i,j} a_{ij}}.$$

The asymptotic distribution of this test statistic is χ^2 with $s - 1$ degrees of freedom. So the p-values can be computed for each SNP. Note that if some of the E_{ij} 's are small, the asymptotic χ^2 distribution may not be appropriate.

1.4 Fisher's exact test

Contrary to the χ^2 test, Fisher's exact test is, as its name states, exact, and it can therefore be used also for small sample sizes. Fisher's test is applied when we have data that are divided into two categories X and Y in two separate ways X_1 and X_2 , and Y_1 and Y_2 , respectively. We denote the row sums by R_i , the column sums by C_j and the entries by a_{ij} :

	Y_1	Y_2	Total
X_1	a_{11}	a_{12}	R_1
X_2	a_{21}	a_{22}	R_2
Total	C_1	C_2	N

Given particular row and column sums the probability of obtaining the above contingency table is given by the *hypergeometric distribution*:

$$\begin{aligned} \mathbb{P}(a_{11}, a_{12}, a_{21}, a_{22}) &= \frac{\binom{R_1}{a_{11}} \binom{R_2}{a_{21}}}{\binom{N}{C_1}} \\ &= \frac{R_1! R_2! C_1! C_2!}{N! a_{11}! a_{12}! a_{21}! a_{22}!}. \end{aligned} \tag{1}$$

Now all possible contingency tables (the entries are nonnegative integers) consistent with the sufficient statistics, namely the row sums R_i and the column sums C_j , have to be found. For each such contingency table the associated conditional probability is calculated using (1). Having to find every contingency table with the given row and column sums makes it computationally expensive with large samples.

To compute the p-value for testing the two variables for independence, the tables must then be brought in order by their χ^2 -scores, or equivalently by their conditional probability, which can be regarded as a measure for dependence. The p-value is then obtained by summing together the probabilities of those tables that represent equal or greater deviation from independence than the observed table, that is those tables with equal or smaller conditional probability than the observed contingency table. We clarify the above explanations by the following example.

Example 1.1. Let X be a marker taking values 0 and 1, and let Y be the disease status, where 0 means non-diseased and 1 diseased. If among the non-diseased individuals five have haplotype 0 and one has haplotype 1, whereas among the diseased individuals none have haplotype 0 and four have haplotype 1, then the contingency table looks as follows:

		Disease status:		Total
		0	1	
Marker:	0	5	0	5
	1	1	4	5
Total		6	4	10

The conditional probability given the row and column sums for this contingency table is

$$\mathbb{P}(5, 0, 1, 4) = \frac{5! 5! 6! 4!}{10! 5! 0! 1! 4!} = 0.0238.$$

Given the row and column sums, the other possible tables and their conditional probabilities are:

$$\begin{aligned} \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix} & \quad \mathbb{P}(4, 1, 2, 3) = 0.2381 \\ \begin{pmatrix} 3 & 2 \\ 3 & 2 \end{pmatrix} & \quad \mathbb{P}(3, 2, 3, 2) = 0.4762 \\ \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix} & \quad \mathbb{P}(2, 3, 4, 1) = 0.2381 \\ \begin{pmatrix} 1 & 4 \\ 5 & 0 \end{pmatrix} & \quad \mathbb{P}(1, 4, 5, 0) = 0.0238 \end{aligned}$$

The resulting p-value is 0.0476. So in this example, there is a statistically significant association between the marker and the disease.

Using the asymptotic χ^2 -test with three degrees of freedom to compute the p-value would not be appropriate in this example: the χ^2 -score for the given contingency table is $\frac{20}{3}$ leading to a p-value of 0.08, which is not significant.

One can imagine that Fisher's exact test can, up to the hard problem of sampling contingency tables with fixed margins, easily be extended to the case of multidimensional contingency tables. In the following subsections we introduce some algebraic tools which enable us to sample contingency tables with fixed margins also if the contingency tables are multidimensional.

1.5 Toric ideals

In this subsection we give the main definitions and some results on toric ideals. However, we do not give any proofs and we assume some basic knowledge of algebra for example on ideals and Groebner bases. Further details on toric ideals including all proofs can be found in Sturmfels (1995).

Toric ideals are a special class of ideals in $k[\mathbf{x}]$. In this chapter we will always work over the complex field \mathbb{C} as some results require that $\text{char}(k) = 0$, that is $|k| = \infty$. We fix a $d \times n$ matrix $\mathcal{A} := (\mathbf{a}_1 \dots \mathbf{a}_n)$ with $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{Z}^d$. Then the group homomorphism defined by the matrix \mathcal{A} is

$$\pi : \mathbb{Z}^n \rightarrow \mathbb{Z}^d, \quad \mathbf{u} = (u_1, \dots, u_n) \mapsto \mathcal{A} \cdot \mathbf{u} = u_1 \mathbf{a}_1 + \dots + u_n \mathbf{a}_n.$$

We can restrict this group homomorphism to a semigroup homomorphism $\pi|_{\mathbb{N}^n} : \mathbb{N}^n \rightarrow \mathbb{Z}^d$. The map $\pi|_{\mathbb{N}^n}$ lifts to a homomorphism of semigroup algebras:

$$\widetilde{\pi|_{\mathbb{N}^n}} : \mathbb{C}[\mathbf{x}] \rightarrow \mathbb{C}[\mathbf{t}^{\pm 1}], \quad x_i \mapsto \mathbf{t}^{\mathbf{a}_i},$$

where $\mathbb{C}[\mathbf{t}^{\pm 1}] := \mathbb{C}[t_1, \dots, t_d, t_1^{-1}, \dots, t_d^{-1}]$ is the Laurent polynomial ring.

Definition 1.2. The kernel of $\widetilde{\pi|_{\mathbb{N}^n}}$ is denoted by $I_{\mathcal{A}}$ and called the **toric ideal** of \mathcal{A} .

Example 1.3. Consider the 3×4 matrix $\mathcal{A} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$. Then

$$\widetilde{\pi|_{\mathbb{N}^4}} : \mathbb{C}[x_1, \dots, x_4] \rightarrow \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}, t_3^{\pm 1}], \quad x_1 \mapsto t_1 t_2, \quad x_2 \mapsto t_1, \quad x_3 \mapsto t_2 t_3, \quad x_4 \mapsto t_3$$

and

$$I_{\mathcal{A}} = \langle x_1 x_4 - x_2 x_3 \rangle.$$

Definition 1.4. Let $f \in \mathbb{C}[\mathbf{x}]$ be a polynomial and $I \subset \mathbb{C}[\mathbf{x}]$ an ideal.

- (i) The polynomial f is called **binomial** if f can be written as difference of two monomials, that is $f = \mathbf{x}^{\mathbf{z}_1} - \mathbf{x}^{\mathbf{z}_2}$, $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{N}^n$.
- (ii) The ideal I is called **binomial ideal** if it is generated by binomials.

Lemma 1.5. *The toric ideal $I_{\mathcal{A}}$ is a binomial ideal:*

$$I_{\mathcal{A}} = \langle \mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{v}} \mid \mathbf{u}, \mathbf{v} \in \mathbb{N}^n \text{ with } \pi(\mathbf{u}) = \pi(\mathbf{v}) \rangle.$$

Every vector $\mathbf{u} \in \mathbb{Z}^n$ can be written uniquely as $\mathbf{u} = \mathbf{u}^+ - \mathbf{u}^-$, where \mathbf{u}^+ and \mathbf{u}^- are non-negative and have disjoint support. More precisely, the i^{th} coordinate of \mathbf{u}^+ equals u_i if $u_i > 0$ and it equals 0 otherwise. With this notation Lemma 1.5 can be rephrased as follows.

Corollary 1.6. $I_{\mathcal{A}} = \langle \mathbf{x}^{\mathbf{u}^+} - \mathbf{x}^{\mathbf{u}^-} \mid \mathbf{u} \in \ker(\pi) \rangle$.

Remark 1.7. It is important to note that in general

$$\ker(\pi) = \langle \mathbf{u}_1, \dots, \mathbf{u}_s \rangle \subset \mathbb{Z}^n \not\Rightarrow I_{\mathcal{A}} = \langle \mathbf{x}^{\mathbf{u}_1^+} - \mathbf{x}^{\mathbf{u}_1^-}, \dots, \mathbf{x}^{\mathbf{u}_s^+} - \mathbf{x}^{\mathbf{u}_s^-} \rangle.$$

This means that it is in general not sufficient to consider only a generating set of the kernel to obtain a generating set for $I_{\mathcal{A}}$. In the following we will explain an algorithm that computes a generating set for the toric ideal to a given matrix \mathcal{A} . The algorithm is based on the following two lemmata. Proofs and further details can be found in Sturmfels (1995).

Let \mathcal{A}_i denote the matrix \mathcal{A} with the column \mathbf{a}_i replaced by its negative $-\mathbf{a}_i$. Let $>_i$ be any term order on $\mathbb{C}[\mathbf{x}]$ such that $x_i >_i x_j$ for all $j \neq i$. In the following lemmata we use $\mathbf{x}^{\mathbf{u}}$, $\mathbf{x}^{\mathbf{v}}$, $\mathbf{x}^{\mathbf{u}^j}$, $\mathbf{x}^{\mathbf{v}^j}$ to denote monomials which do not contain the variable x_i .

Lemma 1.8. *Define $J := \{j \in \{1, \dots, n\} \mid \exists c \in \ker(\pi) \text{ with } c_j \neq 0\}$.*

- (i) $\exists \mathbf{u} \in \ker(\pi)$ such that $u_j \neq 0$ for all $j \in J$.
- (ii) Let C be a generating set for $\ker(\pi)$ such that $\exists \mathbf{u} \in \ker(\pi)$ with $u_j > 0$ for all $j \in J$. Then, $I_{\mathcal{A}} = \langle \mathbf{x}^{\mathbf{v}^+} - \mathbf{x}^{\mathbf{v}^-} \mid \mathbf{v} \in C \rangle$. So in this case it suffices to consider a generating set of the kernel to obtain a generating set of the corresponding toric ideal.

Lemma 1.9. *Let $\mathcal{G}_i = \{x_i^{r_j} \mathbf{x}^{\mathbf{u}^j} - \mathbf{x}^{\mathbf{v}^j} \mid j = 1, \dots, m\}$ be a Groebner basis for $I_{\mathcal{A}_i}$ with respect to $>_i$. Then $\mathcal{G} = \{\mathbf{x}^{\mathbf{u}^j} - x_i^{r_j} \mathbf{x}^{\mathbf{v}^j} \mid j = 1, \dots, m\}$ is a generating set for $I_{\mathcal{A}}$.*

With this preparation it is now easy to construct an algorithm for computing a generating set of a toric ideal.

Proposition 1.10 (Computing a generating set of a toric ideal). *Let \mathcal{A} be a $d \times n$ matrix with entries in \mathbb{Z} . Then a generating set for the toric ideal $I_{\mathcal{A}}$ can be constructed in a finite number of steps by the following algorithm:*

Input: \mathcal{A}

Output: a generating set for $I_{\mathcal{A}}$

1. Choose a subset $\{i_1, \dots, i_r\} \subset \{1, \dots, n\}$ such that $\exists \mathbf{u} \in \ker(\mathcal{A}_{i_1 i_2 \dots i_r})$ with $u_j > 0$ for all $j \in J$, where $\mathcal{A}_{i_1 \dots i_r} := (\dots (\mathcal{A}_{i_1})_{i_2} \dots)_{i_r}$ and J is defined as in Lemma 1.8.
2. Find a basis C for the kernel of $\mathcal{A}_{i_1 i_2 \dots i_r}$ such that C contains a vector \mathbf{v} with $v_j > 0$ for all $j \in J$.
3. $I_{\mathcal{A}_{i_1 i_2 \dots i_r}} = \langle \mathbf{x}^{\mathbf{u}^+} - \mathbf{x}^{\mathbf{u}^-} \mid u \in C \rangle$ (by Lemma 1.8 (ii)).
4. Let $l := r$.
5. While $l \geq 1$ do:
 - 5.1. Choose a monomial ordering $>_{i_l}$.
 - 5.2. Compute a Groebner basis $\mathcal{G}_{i_1 i_2 \dots i_l}$ for $I_{\mathcal{A}_{i_1 i_2 \dots i_l}}$.
 - 5.3. Flip the variable x_{i_l} as in Lemma 1.9 to get generators for $I_{\mathcal{A}_{i_1 i_2 \dots i_{l-1}}}$.
 - 5.4. $l := l - 1$
6. Output the resulting generating set for $I_{\mathcal{A}}$.

We illustrate this algorithm by the following example.

Example 1.11. The 2×4 matrix

$$\mathcal{A} := \begin{pmatrix} 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix}$$

is given. First we need to compute the kernel of \mathcal{A} :

$$\ker(\mathcal{A}) = \left\langle \left(\begin{array}{c} 1 \\ -2 \\ 1 \\ 0 \end{array} \right), \left(\begin{array}{c} 2 \\ -3 \\ 0 \\ 1 \end{array} \right) \right\rangle.$$

The kernel contains the vector $\begin{pmatrix} 3 \\ -5 \\ 1 \\ 1 \end{pmatrix}$, which is non-zero in any component. As the second component is negative, we replace the second column of \mathcal{A} by its negative. So

$$\mathcal{A}_2 = \begin{pmatrix} 3 & -2 & 1 & 0 \\ 0 & -1 & 2 & 3 \end{pmatrix}$$

and the kernel of \mathcal{A}_2 is

$$\ker(\mathcal{A}) = \left\langle \left(\begin{array}{c} 1 \\ 2 \\ 1 \\ 0 \end{array} \right), \left(\begin{array}{c} 2 \\ 3 \\ 0 \\ 1 \end{array} \right) \right\rangle = \left\langle \left(\begin{array}{c} 1 \\ 2 \\ 1 \\ 0 \end{array} \right), \left(\begin{array}{c} 3 \\ 5 \\ 1 \\ 1 \end{array} \right) \right\rangle =: \langle C \rangle.$$

So we have found C and can now compute a generating set for $I_{\mathcal{A}_2}$:

$$I_{\mathcal{A}_2} = \langle x_1 x_2^2 x_3 - 1, x_1^3 x_2^5 x_3 x_4 - 1 \rangle.$$

We choose the monomial ordering $>_2$ to be $x_2 >_2 x_3 >_2 x_4 >_2 x_1$. With the computer algebra system CoCoA (<http://cocoa.dima.unige.it/>) we compute a Groebner basis for $I_{\mathcal{A}_2}$. The resulting Groebner basis is:

$$I_{\mathcal{A}_2} = \langle x_1 x_2^2 x_3 - 1, x_1 x_2 x_4 - x_3, x_2 x_3^2 - x_4, x_3^3 - x_1 x_4^2 \rangle.$$

By flipping the variable x_2 we get a generating set for $I_{\mathcal{A}}$:

$$I_{\mathcal{A}} = \langle x_1 x_3 - x_2^2, x_1 x_4 - x_2 x_3, x_3^2 - x_2 x_4, x_3^3 - x_1 x_4^2 \rangle.$$

This generating set is not necessarily minimal or even unique. With the help of CoCoA we can compute the unique reduced Groebner basis of this ideal with respect to the lexicographic order:

$$I_{\mathcal{A}} = \langle x_2 x_4 - x_3^2, x_1 x_4 - x_2 x_3, x_1 x_3 - x_2^2 \rangle.$$

The corresponding projective toric variety, that is the variety generated by this toric ideal, is the twisted cubic curve in \mathbb{P}^3 , well-known in algebraic geometry.

Until now, we have presented some results on toric ideals and an algorithm using Groebner bases. This algorithm gives an idea of the importance of Groebner bases as algorithmic tools for algebraic computations. However, we did not yet build a connection between toric ideals and contingency tables, Fisher's exact test, or statistics in general. This is the aim of the next subsection.

1.6 Sampling on the space of contingency tables with fixed margins

This subsection explains how the theory of toric ideals can be used to sample contingency tables with fixed margins, even if the contingency tables are multidimensional.

In order to give an idea of the main concepts of this application of toric ideals, we start just with two-dimensional tables, more precisely $d_1 \times d_2$ contingency tables, and generalize these concepts to multidimensional tables only at the end, when we get to the main result.

In what follows, we will switch back and forth from the algebraic language to the language of contingency tables.

Notation 1.12. We fix d_1 and $d_2 \in \mathbb{N}$ and look at $d_1 \times d_2$ contingency tables.

(i) Let $\mathbf{t} = (t_1, \dots, t_{d_1}, t_{d_1+1}, \dots, t_{d_1+d_2})^T$ be a vector in $\mathbb{N}^{d_1+d_2}$ such that

$$\sum_{i=1}^{d_1} t_i = \sum_{j=d_1+1}^{d_1+d_2} t_j.$$

This means that \mathbf{t} is a possible vector containing the row and column sums of a $d_1 \times d_2$ contingency table.

(ii) The map T is defined as follows:

$$T : I_1 \times I_2 \rightarrow \mathbb{N}^{d_1+d_2}, \quad (i, j) \mapsto (0, 0, \dots, 0, 1_i, 0, \dots, 0, 1_{d_1+j}, 0, \dots, 0)^T,$$

where $I_1 := \{1, \dots, d_1\}$ and $I_2 := \{1, \dots, d_2\}$.

(iii) \mathcal{A} is a $(d_1 + d_2) \times (d_1 \cdot d_2)$ matrix whose columns are given by the image of T :

$$\mathcal{A} = \left(\begin{array}{ccc|ccc|cc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{array} \right).$$

(iv) For a given vector $\mathbf{t} \in \mathbb{N}^{d_1+d_2}$ with the properties stated above the set $\mathcal{F}_{\mathbf{t}}$ is defined as follows:

$$\mathcal{F}_{\mathbf{t}} := \left\{ f : I_1 \times I_2 \rightarrow \mathbb{N} \mid \sum_{(i,j) \in I_1 \times I_2} f(i,j) \cdot T(i,j) = \mathbf{t} \right\}.$$

This set corresponds to the set of all contingency tables with the row and column sums given by the vector \mathbf{t} .

Example 1.13. We look again at the contingency table of Example 1.1. In this case,

$$\mathbf{t} = \begin{pmatrix} 5 \\ 5 \\ 6 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathcal{A} = \left(\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \hline 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array} \right).$$

It is important to note that if we write the contingency table of Example 1.1 as a vector in the following way

$$\mathbf{u} = \begin{pmatrix} 5 \\ 0 \\ 1 \\ 4 \end{pmatrix},$$

then

$$\mathcal{A} \cdot \mathbf{u} = \mathbf{t}.$$

This example leads us to the following observation building a connection between contingency tables and toric ideals.

Remark 1.14. We have seen in the above example that $\mathcal{A} \cdot \mathbf{u} = \mathbf{t}$. It is clear that this equation is true in general and not only for this example. So we obtain

$$\begin{aligned} \ker(\mathcal{A}) &= \{\text{tables with all row and column sums equal to zero}\} \\ &= \{f : I_1 \times I_2 \rightarrow \mathbb{Z} \mid \sum_{(i,j) \in I_1 \times I_2} f(i,j) \cdot T(i,j) = 0\}, \end{aligned}$$

and therefore

$$\begin{aligned} I_{\mathcal{A}} &= \langle \mathbf{x}^{\mathbf{u}^+} - \mathbf{x}^{\mathbf{u}^-} \mid \mathbf{u} \in \ker(\mathcal{A}) \rangle \\ &= \langle \mathbf{x}^{f^+} - \mathbf{x}^{f^-} \mid f : I_1 \times I_2 \rightarrow \mathbb{Z}, \sum_{(i,j) \in I_1 \times I_2} f(i,j) \cdot T(i,j) = 0 \rangle, \end{aligned}$$

where $f^+(i,j) = \max(f(i,j), 0)$ and analogously $f^-(i,j) = \max(-f(i,j), 0)$.

We now switch to the case of multidimensional contingency tables. This means that we fix $d_1, \dots, d_n \in \mathbb{N}$ and define $I_1 := \{1, \dots, d_1\}, \dots, I_n := \{1, \dots, d_n\}$. The vector \mathbf{t} , the map T , the matrix \mathcal{A} and the set $\mathcal{F}_{\mathbf{t}}$ can be generalized to the multidimensional case. We will illustrate this with an example in Section 2.1.

The notion of a Markov basis will be very important for the following parts of this paper.

Definition 1.15. A **Markov basis** (to a given matrix of the form \mathcal{A}) is a finite family of functions $(f_1, \dots, f_L : I_1 \times \dots \times I_n \rightarrow \mathbb{Z})$ such that

$$(i) \quad \sum_{(i_1, \dots, i_n) \in I_1 \times \dots \times I_n} f_l(i_1, \dots, i_n) \cdot T(i_1, \dots, i_n) = 0 \quad \text{for } 1 \leq l \leq L, \quad (2)$$

(ii) for any \mathbf{t} and $f, f' \in \mathcal{F}_{\mathbf{t}}$ there are $(\varepsilon_1, f_{i_1}), \dots, (\varepsilon_A, f_{i_A})$ with $\varepsilon \in \{\pm 1\}$,

$$f' = f + \sum_{j=1}^A \varepsilon_j f_{l_j} \quad \text{and} \quad f + \sum_{j=1}^a \varepsilon_j f_{l_j} \geq 0 \quad \text{for } 1 \leq a \leq A. \quad (3)$$

A Markov basis allows the construction of a Markov chain on $\mathcal{F}_{\mathbf{t}}$, that is the space of all contingency tables with the margins given by the vector \mathbf{t} , as described and proven in Diaconis et al. (1998) as follows:

Proposition 1.16. *Let $\sigma(g)$ ¹ be a positive function on $\mathcal{F}_{\mathbf{t}}$. Given a Markov basis f_1, \dots, f_L , start with an arbitrary state $f \in \mathcal{F}_{\mathbf{t}}$ and generate a Markov chain on $\mathcal{F}_{\mathbf{t}}$ by choosing l uniformly in $1, \dots, L$ and ε in $\{\pm 1\}$ independent of l . If the chain is currently at state $g \in \mathcal{F}_{\mathbf{t}}$, compute the proposed state $g + \varepsilon \cdot f_l$. If $g + \varepsilon \cdot f_l$ is nonnegative, move to the new proposed state with probability*

$$P(\text{move}) = \min \left\{ 1, \frac{\sigma(g + \varepsilon f_l)}{\sigma(g)} \right\}.$$

Otherwise the chain remains at g . This is an irreducible, aperiodic and positive recurrent Markov chain on $\mathcal{F}_{\mathbf{t}}$ with stationary distribution proportional to $\sigma(g)$.

This proposition shows that if we have a Markov basis it is possible to generate a Markov chain on the space of contingency tables with fixed margins. So the problem of sampling contingency tables with fixed margins is reduced to the problem of finding a Markov basis to a given matrix \mathcal{A} . For 2×2 contingency tables the Markov basis trivially consists of the tables $\pm \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$. However, for multidimensional contingency tables finding a Markov basis is not trivial. But the following theorem that is proven in Diaconis et al. (1998) enables us to easily find a Markov basis also for multidimensional contingency tables.

Theorem 1.17. *A finite family of functions $(f_1, \dots, f_L : I_1 \times \dots \times I_n \rightarrow \mathbb{Z})$ is a Markov basis (to a given matrix of the form \mathcal{A}) if and only if*

$$\langle \mathbf{x}^{f_1^+} - \mathbf{x}^{f_1^-}, \dots, \mathbf{x}^{f_L^+} - \mathbf{x}^{f_L^-} \rangle = I_{\mathcal{A}}.$$

So the problem of finding a Markov basis to a given matrix \mathcal{A} is reduced to the problem of finding a generating set of the toric ideal defined by the matrix \mathcal{A} , which can be solved by the algorithm presented in Section 1.5. In the following subsection we show how these concepts can be used to generalize Fisher's exact test to multidimensional tables.

1.7 Extended Fisher's exact test

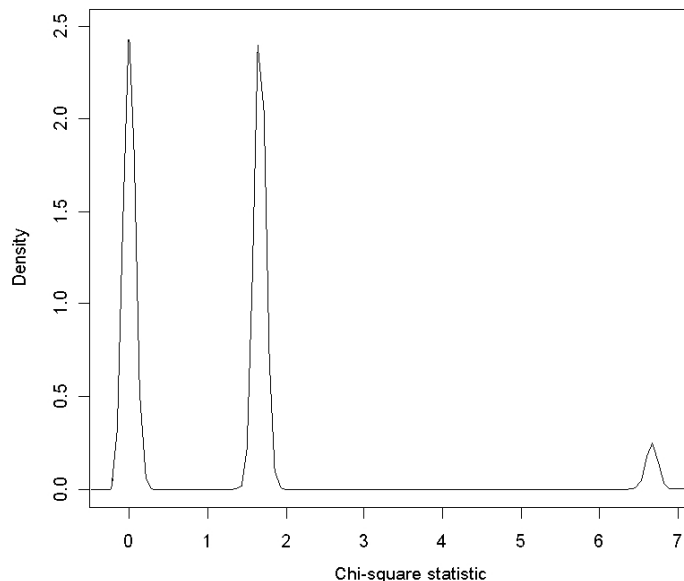
For two-dimensional tables with small counts it is possible to find all contingency tables with given row and column sums and to compute the exact p-value by Fisher's exact test as presented in Section 1.4. However, this gets difficult for tables with large

¹In what follows $\sigma(g)$ is the hypergeometric distribution.

counts or for more-dimensional tables. The χ^2 -statistic might not be well approximated by the χ^2 -distribution. So we would like to use the MCMC approach discussed above to sample from the space of all contingency tables with given row and column sums and approximate the exact p-value by the quantiles of the resulting posterior distribution. So for every sampled contingency table we will compute its χ^2 -statistic and then look at the posterior distribution of this statistic.

Going back to our Example 1.1, we are given a 2×2 table. So the Markov basis consists only of the elements $\pm \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$. The resulting posterior density of the MCMC approach described above is shown in Figure 1. The 0.95-quantile is 1.667. So there is a significant association in the given table, which has a χ^2 -statistic of 6.667.

Figure 1: Posterior density of the χ^2 -statistic.



We will now turn to our model and explain how all these concepts relate to finding interactions between loci in disease association studies.

2 Model

In this section we will first describe the general case of testing for no two-way interactions in multidimensional contingency tables. We will show that when testing for no two-way interactions the Markov basis only consists of elements of the form $\pm \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$ for contingency tables of any dimension. We will then shortly present

how the matrix \mathcal{A} is computed for testing for no k -way interactions in $I_1 \times \cdots \times I_n$ contingency tables. Further, we discuss various interaction tests that are of interest for disease association studies and give the sufficient statistics which are needed to compute the matrix \mathcal{A} and perform the extended Fisher's exact test. Finally, we shortly review the whole procedure for testing for interactions within multidimensional contingency tables.

2.1 Testing for no two-way interactions

In this subsection we construct the extended Fisher's exact test for testing the null hypothesis of no two-way interactions. We will look at an example where we want to test four SNPs for independence. Every marker can take on the values 0 and 1. So the data can be gathered in a $2 \times 2 \times 2 \times 2$ contingency table.

First, we want to find a Markov basis for testing for no two-way interactions in $2 \times 2 \times 2 \times 2$ contingency tables. For this, we need to construct the corresponding matrix \mathcal{A} and find a generating set for the toric ideal $I_{\mathcal{A}}$.

For $2 \times 2 \times 2 \times 2$ contingency tables $I_1 = I_2 = I_3 = I_4 = \{1, 2\}$ and the vector \mathbf{t} is of the form

$$\mathbf{t} = (t_1, t_2, \dots, t_8)^T \in \mathbb{N}^8 \quad \text{such that} \quad t_1 + t_2 = t_3 + t_4 = t_5 + t_6 = t_7 + t_8.$$

This vector represents the 'row and column' sums. For n -dimensional tables these are the sums of the entries of the $(n - 1)$ -dimensional subtables. The map T for $2 \times 2 \times 2 \times 2$ contingency tables is

$$T : I_1 \times I_2 \times I_3 \times I_4 \rightarrow \mathbb{N}^8, \quad \text{where for example } (1, 2, 2, 1) \mapsto (1, 0, 0, 1, 0, 1, 1, 0)^T$$

and the corresponding matrix \mathcal{A} is the 8×16 matrix

$$\mathcal{A} = \left(\begin{array}{cccc|cccc|cccc|cccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ \hline 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{array} \right).$$

With the algorithm given in Proposition 1.10 we can compute a generating set for $I_{\mathcal{A}}$:

$$I_{\mathcal{A}} = \langle x_3x_6 - x_4x_5, x_3x_{13} - x_7x_9, x_1x_4 - x_2x_3, x_1x_7 - x_3x_5, x_1x_{15} - x_7x_9, x_{11}x_{14} - x_{12}x_{13}, \\ x_1x_{14} - x_6x_9, x_1x_{13} - x_5x_9, x_1x_{12} - x_4x_9, x_3x_{15} - x_7x_{11}, x_3x_8 - x_4x_7, \\ x_5x_8 - x_6x_7, x_1x_8 - x_4x_5, x_2x_8 - x_4x_6, x_7x_{10} - x_8x_9, x_1x_{16} - x_8x_9, x_7x_{12} - x_8x_{11}, \\ x_4x_{16} - x_8x_{12}, x_7x_{14} - x_8x_{13}, x_6x_{16} - x_8x_{14}, x_7x_{16} - x_8x_{15}, x_3x_{10} - x_4x_9, \\ x_{10}x_{15} - x_{12}x_{13}, x_5x_{10} - x_6x_9, x_9x_{12} - x_{10}x_{11}, x_6x_{12} - x_8x_{10}, x_2x_{14} - x_6x_{10}, \\ x_{10}x_{16} - x_{12}x_{14}, x_2x_{12} - x_4x_{10}, x_1x_{10} - x_2x_9, x_9x_{14} - x_{10}x_{13}, x_{11}x_{16} - x_{12}x_{15}, \\ x_9x_{16} - x_{12}x_{13}, x_9x_{15} - x_{11}x_{13}, x_6x_{11} - x_8x_9, x_5x_{11} - x_7x_9, x_2x_{11} - x_4x_9, \\ x_3x_{12} - x_4x_{11}, x_1x_{11} - x_3x_9, x_5x_{16} - x_8x_{13}, x_5x_{15} - x_7x_{13}, x_{13}x_{16} - x_{14}x_{15}, \\ x_6x_{15} - x_8x_{13}, x_5x_{14} - x_6x_{13}, x_2x_7 - x_4x_5, x_2x_{16} - x_8x_{10}, x_4x_{13} - x_8x_9, \\ x_2x_{15} - x_8x_9, x_2x_{13} - x_6x_9, x_1x_6 - x_2x_5, x_4x_{14} - x_8x_{10}, x_5x_{12} - x_8x_9, \\ x_3x_{16} - x_8x_{11}, x_3x_{14} - x_8x_9, x_4x_{15} - x_8x_{11} \rangle.$$

Each of these generators corresponds to an element of the Markov basis and at the same time, it represents a table. We denote a $2 \times 2 \times 2 \times 2$ table by y and the cells by $y[i, j, k, l]$, where i denotes the first marker, j the second marker, k the third marker and l the fourth marker, $i, j, k, l \in \{0, 1\}$. We defined the matrix \mathcal{A} such that the following pairs correspond:

$$\begin{array}{cccc} y[0, 0, 0, 0] = x_1 & y[0, 0, 0, 1] = x_2 & y[0, 0, 1, 0] = x_3 & y[0, 0, 1, 1] = x_4 \\ y[0, 1, 0, 0] = x_5 & y[0, 1, 0, 1] = x_6 & y[0, 1, 1, 0] = x_7 & y[0, 1, 1, 1] = x_8 \\ y[1, 0, 0, 0] = x_9 & y[1, 0, 0, 1] = x_{10} & y[1, 0, 1, 0] = x_{11} & y[1, 0, 1, 1] = x_{12} \\ y[1, 1, 0, 0] = x_{13} & y[1, 1, 0, 1] = x_{14} & y[1, 1, 1, 0] = x_{15} & y[1, 1, 1, 1] = x_{16}. \end{array}$$

So the generator $x_3x_6 - x_4x_5$ corresponds to the table with the cell entries $y[0, 0, 1, 0] = y[0, 1, 0, 1] = 1$, $y[0, 0, 1, 1] = y[0, 1, 0, 0] = -1$ and zero entries in the remaining cells.

As the Markov basis only consists of tables of the form $\pm \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$, the acceptance probability in the MCMC algorithm for performing the extended Fisher's exact test gets very simple: It is just the quotient of the four entries that change from the old to the new contingency table. So for example, if the chain currently is in contingency table a , the element $x_3x_6 - x_4x_5$ is chosen from the Markov basis, $\epsilon = 1$ and all the new entries are positive, then the acceptance probability is

$$P(\text{move}) = \min \left(1, \frac{a_{0010}! a_{0101}! a_{0011}! a_{0100}!}{(a_{0010} + 1)! (a_{0101} + 1)! (a_{0011} - 1)! (a_{0100} - 1)!} \right).$$

It is clear that we can find a Markov basis for $I \times J$ contingency tables consisting only of tables of the form $\pm \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$, which means that the reduced Groebner

basis of the corresponding toric ideal consists only of binomials of degree 2 of the form $y_i y_j - y_k y_l$ with $i \neq j$ and $k \neq l$. However, it might seem surprising that the reduced Groebner basis of the toric ideal corresponding to $2 \times 2 \times 2 \times 2$ contingency tables also only consists of binomials of this particular form. This means that we can find a minimal Markov basis consisting merely of tables of the form $\pm \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$.

But notice that the matrix \mathcal{A} given above describes the Segre variety

$$\Sigma_{2,2,2,2} := \{\langle y_{ijkl} \rangle \subset \mathbb{P}^{15} \mid \text{rg}[y_{ijkl}] = 1\}.$$

So the toric ideal $I_{\mathcal{A}}$ is just the vanishing ideal of this Segre variety. In order to compute the vanishing ideal of this variety, we will review some definitions and results on n^{th} order tensors. Further details including all the proofs can be found in Ha (2002).

Definition 2.1. Let \mathcal{P} be a generic n^{th} order tensor.

i) For each $l \in \{1, \dots, n\}$,

$$p_{i_1 \dots i_l \dots i_n} p_{j_1 \dots j_l \dots j_n} - p_{i_1 \dots i_{l-1} j_l i_{l+1} \dots i_n} p_{j_1 \dots j_{l-1} i_l j_{l+1} \dots j_n} \in \mathbb{C}[\mathcal{P}]$$

is called a **2×2 minor about the l^{th} coordinate of \mathcal{P}** .

ii) A **2×2 minor** is a 2×2 minor about at least one of its coordinates.

iii) $I_2(\mathcal{P}) := \langle 2 \times 2 \text{ minors of } \mathcal{P} \rangle \subset \mathbb{C}[\mathcal{P}]$

Theorem 2.2. i) $I_2(\mathcal{P})$ describes the tensors of rank 1.

ii) $I_2(\mathcal{P}) \subset \mathbb{C}[\mathcal{P}]$ is a prime ideal.

Corollary 2.3. Consider the Segre embedding $\sigma : \mathbb{P}^{I_1-1} \times \dots \times \mathbb{P}^{I_n-1} \rightarrow \mathbb{P}^{I_1 \dots I_n-1}$. Then $J(\Sigma_{I_1, \dots, I_n}) = I_2(\mathcal{P})$, where \mathcal{P} is a generic $I_1 \times \dots \times I_n$ tensor.

So from this corollary it is clear that the toric ideal $I_{\mathcal{A}}$ defined above is generated by binomials of degree 2. Moreover, this corollary tells us that when testing for 2-way interactions in any contingency table the reduced Groebner basis of the corresponding toric ideal consists merely of binomials of the form $x_i x_j - x_k x_l$ with $i \neq j$ and $k \neq l$, independently of the chosen monomial ordering. This result is stated in the following proposition.

Proposition 2.4. When testing for two-way interactions in contingency tables of an arbitrary dimension, there exists a Markov basis consisting of elements of the form $\pm \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ only.

2.2 Computing the matrix \mathcal{A}

In this subsection we will discuss how to compute the matrix \mathcal{A} when testing for no k -way interactions in a $I_1 \times \cdots \times I_n$ contingency table.

It is known and further discussed in Section 2.3 that the sufficient statistics for testing for k -way interactions are the $k - 1$ dimensional marginal tables resulting by summing over $n - k + 1$ indices. For simplicity of the following discussion lets assume that $I_1 = \dots = I_n = N$. Then \mathcal{A} is a $\binom{n}{n-k+1} N^{k-1} \times N^n$ -matrix where each row represents a cell of a marginal subtable and each column a cell of the contingency table. The entry $\mathcal{A}[i, j]$ is 1 if the corresponding cell of the contingency table has been summed over in the corresponding cell of the marginal subtable. Otherwise the entry is 0. This is illustrated with the following example where $k = n = 3$, $N = 2$:

$$\mathcal{A} = \begin{array}{c|cccccccc} & 111 & 112 & 121 & 122 & 211 & 212 & 221 & 222 \\ \hline 11\cdot & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12\cdot & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 21\cdot & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 22\cdot & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 1\cdot 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1\cdot 2 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2\cdot 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 2\cdot 2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ \hline \cdot 11 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \cdot 12 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ \cdot 21 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ \cdot 22 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}$$

Note that some nice symmetry arguments can be used to compute this matrix, without needing to go over every cell entry. We first build an auxiliary matrix of dimension $\binom{n}{k-1} \times n$, where the rows hold all combinations of $k - 1$ ones and $n - k + 1$ zeros. Here 0 means summing over that index. So the auxiliary matrix for the example above is:

$$M = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

We can get \mathcal{A} from M by going over the matrix M row by row starting from the last entry. A 0 in M means copying the preceding submatrix to the right of it, a 1 means copying the preceding submatrix to the right and beneath it. This procedure is best explained by looking at the example above.

2.3 Various interaction tests

In this subsection we present various hypotheses that can easily be tested with the extended Fisher's exact test and discuss some hypotheses that are particularly interesting for disease association studies. For simplicity we constrain this discussion to the case of three variables, namely two markers X and Y and the disease state D . All proofs and further explanations can be found in Bishop et al. (1975).

For three variables we can define the models given in Table 1 and then compute their fit to the data by using the extended Fisher's exact test. We use the notation presented in Bishop et al. (1975) to denote the different models. The model assumes interaction between the variables listed in the model and tests for all combinations that are not listed. So the model (X, Y, Z) in the table below represents the independence model, the model (XY, XD, YD) the no tree-way interaction model and the other models are intermediate models. The model (XY, Z) for example assumes that there can be an interaction between the two markers and we are testing if there is any association between these markers and the disease. Finally, the model (XD, YD) represents the model of conditional independence of X and Y given D .

Table 1: Interaction models for three-dimensional contingency tables.

Model	Minimal sufficient statistics	Expected counts
(X, Y, D)	$(n_{i..}), (n_{.j.}), (n_{..k})$	$\hat{n}_{ijk} = \frac{n_{i..}n_{.j.}n_{..k}}{(n_{...})^2}$
(XY, D)	$(n_{ij.}), (n_{..k})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{..k}}{(n_{...})}$
(XD, Y)	$(n_{i.k}), (n_{.j.})$	$\hat{n}_{ijk} = \frac{n_{i.k}n_{.j.}}{(n_{...})}$
(X, YD)	$(n_{i..}), (n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{.jk}n_{i..}}{(n_{...})}$
(XY, YD)	$(n_{ij.}), (n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{.jk}}{(n_{.j.})}$
(XY, XD)	$(n_{ij.}), (n_{i.k})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{i.k}}{(n_{i..})}$
(XD, YD)	$(n_{i.k}), (n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{i.k}n_{.jk}}{(n_{..k})}$
(XY, XD, YD)	$(n_{ij.}), (n_{i.k}), (n_{.jk})$	Iterative proportional fitting

Performing the extended Fisher's exact test involves sampling from the space of contingency tables with fixed minimal sufficient statistics and computing the χ^2 -statistic. So the minimal sufficient statistics and the expected counts for each cell of the table need to be calculated. These are given in Table 1. For example for the independence model (X, Y, D) , \hat{n}_{ijk} is computed as follows:

$$\begin{aligned}
 \hat{n}_{ijk} &= n_{...}\hat{p}_{ijk} \\
 &= n_{...}\hat{p}_{i..}\hat{p}_{.j.}\hat{p}_{..k} \\
 &= n_{...} \frac{n_{i..}}{n_{...}} \frac{n_{.j.}}{n_{...}} \frac{n_{..k}}{n_{...}} \\
 &= \frac{n_{i..}n_{.j.}n_{..k}}{(n_{...})^2}.
 \end{aligned}$$

The other counts are computed analogously. Bishop et al. (1975) show that the cell counts cannot directly be estimated when there is a closed loop in the model configuration as for example in (XY, YD, DX) . But in this case, estimates can be achieved by iterative proportional fitting, which is also discussed in Bishop et al. (1975).

For disease association studies two of the above models are particularly interesting, namely the model (XY, D) and the model (XY, XD, YD) . The former model tests for interactions between any of the markers and the disease and the latter tests for interactions between the two markers regarding the disease. We will use these models in the simulation studies presented in Section 3.

When looking at three or more markers one can perform many other interesting tests. For example with four markers W, X, Y and Z we could be interested in the models $M_1 = (WXYZ, D)$, $M_2 = (WXYZ, WD, XD, YD, ZD)$, $M_3 = (WXYZ, WXD, WYD, WZD, XYD, XZD, YZD)$, and $M_4 = (WXYZ, WXYD, WXZD, WYZD, XYZD)$. Note that by the hierarchy principle these models are nested. So a significant p-value in M_4 should lead to a significant value in all the other models.

It is important to note that these tests for interaction between markers necessarily imply working with multidimensional contingency tables and cannot be performed by collapsing the multidimensional tables to 2-dimensional haplotype tables. For example in the case of two markers X and Y , it could be thought that testing for association in Table 2 is the same as testing model (XY, XD, YD) . However, this is not true. The former tests for association between the haplotypes and the disease, whereas the latter tests for interactions between the two markers regarding the disease. The sufficient statistics for the model described in Table 2 are the row and column sums $(n_{i.j.})$ and $(n_{..k})$. So testing for association in this collapsed table is the same as testing the model (XY, D) , which does not test for interactions between markers. So in order to test for interactions between markers, it is inevitable to work with multidimensional tables.

Table 2: Testing for association between haplotypes and disease.

		Disease status:		Total:
		0	1	
Haplotype:	00	n_{000}	n_{001}	$n_{00.}$
	01	n_{010}	n_{100}	$n_{10.}$
	10	n_{100}	n_{101}	$n_{10.}$
	11	n_{110}	n_{111}	$n_{11.}$
Total:		$n_{..0}$	$n_{..1}$	$n_{...}$

2.4 Procedure for performing the extended Fisher’s exact test

By putting together the various pieces described in the previous sections, we are now able to perform Fisher’s exact test for testing for interactions on multidimensional tables. Summarized, the whole procedure works as follows:

- i) Choose a model and determine its minimal sufficient statistics and the expected counts as described in Section 2.3.
- ii) Compute the corresponding matrix \mathcal{A} as described in Section 2.2.
- iii) Compute the corresponding toric ideal $I_{\mathcal{A}}$ as described in Section 1.5.
- iv) Run MCMC starting with a given data matrix and the Markov basis from iii) as described in Section 1.6.
- v) Analyze the posterior distribution of the χ^2 -statistic: the exact p-value of the given data can be approximated by the quantiles of this distribution.

3 Simulations and results

After presenting an algorithm for testing for interactions between markers in multidimensional tables, we will now discuss how this method performs. We simulated data under different multilocus models of disease using the software HAP-SAMPLE (Wright et al. (2007)). In a first stage, we filtered the data with a single locus approach and then used the algorithm above to test for interactions between the disease and two markers chosen by the single locus approach.

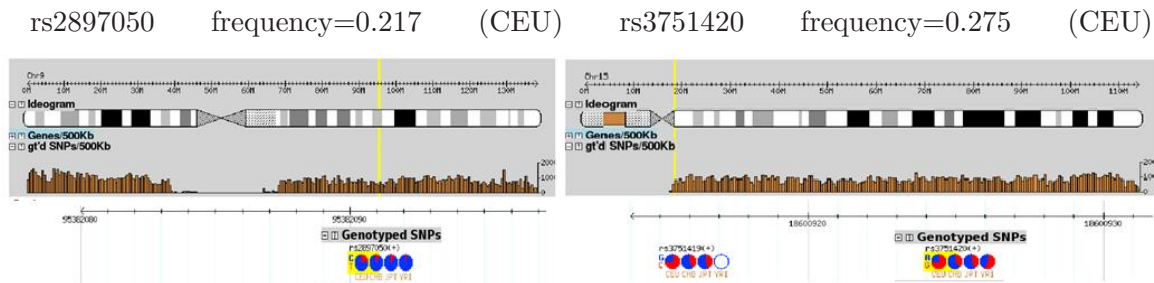
3.1 Simulations with HAP-SAMPLE and disease models

HAP-SAMPLE is a simulator whose main advantage is that it simulates data with realistic allele frequencies and realistic linkage disequilibrium patterns. HAP-SAMPLE resamples SNPs from the HapMap project (The International HapMap Consortium (2003)). The data is obtained by resampling within a population while allowing for recombination in order to mimic the meiosis process. The recombination process respects the linkage disequilibrium structure existing in the original HapMap data. HAP-SAMPLE is based on phase I/II of the HapMap project, which represents more than 3 million SNPs and is thought to contain 25% to 35% of the common genetic variation in the four populations represented.

Another advantage of HAP-SAMPLE is the flexibility in terms of disease models. Indeed, it allows for multiple loci to be causative for the disease. Given the disease status D the probability of a particular genotype g (for one or more loci) can be computed by the probability distribution $p(g|D)$, which can be specified by the user.

For our simulations we used the CEPH population (Utah residents with ancestry from northern and western Europe). We wanted to restrict the dataset to 10'000 SNPs in the interest of time. Therefore, we chose chromosome 9 and 13, which have about 10'000 SNPs in the database. Finally, we chose the two SNPs shown in Figure 2 to be causative and simulated 400 cases and 400 controls.

Figure 2: Showing the causative SNPs chosen on chromosome 9 and 13. Note that our chosen SNP on chromosome 13 is only 250 bp apart from another unrelated SNP.



We then formulated three different disease models. From now on we will focus on genotype data. As described in Section 1.1, the genotype is the sum of the states on each chromosome. For example, if an individual has state 0 on the paternal chromosome and state 1 on the maternal chromosome at a specific locus, then the genotype at that locus is 1.

The first model (M_0) described in Table 3 is a control model, where the probability of being diseased is the same independently of the genotype.

Table 3: Control model (M_0).

		Marker 2:		
		0	1	2
Marker 1:	0	α	α	α
	1	α	α	α
	2	α	α	α

The second model (M_1) described in Table 4 is an additive model, where two markers are responsible for the disease, each of them independently of the other (no epistasis). One can imagine a biochemical metabolic pathway with multiple paths leading to the same product. The two markers might lie in the coding region for two enzymes necessary for these two alternative paths. So mutating two enzymes on two alternative paths will decrease the product in an additive fashion. This is illustrated in Figure 3.

The last model (M_2) is described in Table 5 and has an interaction parameter θ_3 . A possible biological scenario would be having two causative markers in two subunits

Table 4: Additive model, M_1 .

		Marker 2:		
		0	1	2
Marker 1:	0	0	$(1 + \theta_2)$	$(1 + \theta_2)^2$
	1	$(1 + \theta_1)$	$(1 + \theta_1)(1 + \theta_2)$	$(1 + \theta_1)(1 + \theta_2)^2$
	2	$(1 + \theta_1)^2$	$(1 + \theta_1)^2(1 + \theta_2)$	$(1 + \theta_1)^2(1 + \theta_2)^2$

Figure 3: Showing three alternative paths to get a single product P . Each blue dot represents an intermediate molecule. The red marks represent mutated enzymes on two different metabolic reactions. Each mutation will hence decrease the amount of product P by disrupting one path.

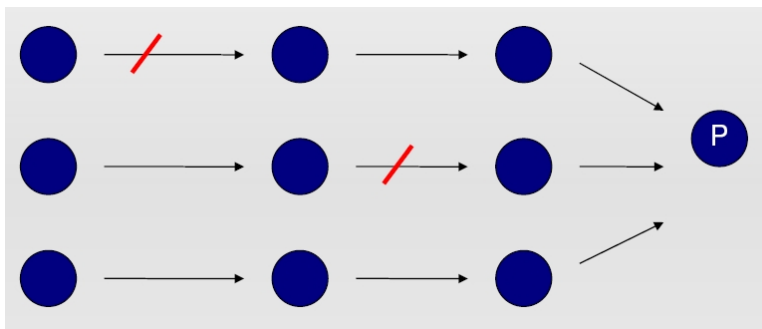


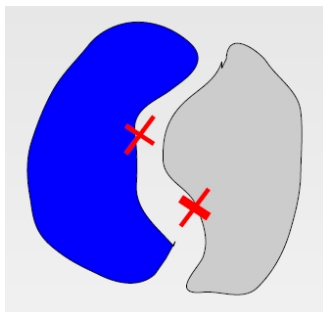
Table 5: Multiplicative model, M_2

		Marker 2:		
		0	1	2
Marker 1:	0	0	$(1 + \theta_2)$	$(1 + \theta_2)^2$
	1	$(1 + \theta_1)$	$(1 + \theta_1)(1 + \theta_2)(1 + \theta_3)$	$(1 + \theta_1)(1 + \theta_2)^2(1 + \theta_3)^2$
	2	$(1 + \theta_2)^2$	$(1 + \theta_1)^2(1 + \theta_2)(1 + \theta_3)^2$	$(1 + \theta_1)^2(1 + \theta_2)^2(1 + \theta_3)^4$

of one complex, e.g. an essential enzyme. Mutating one of the subunits might not alter the structure of the complex enough to disrupt the function, while a mutation in both subunits might. This is described in Figure 4. In this case we have epistasis as the effect of one gene is modified by the other.

As a first step awaiting for further work, we fixed the parameters of the disease models to be $\alpha = 0.5$ for M_0 , $\theta_1 = 0.5$, $\theta_2 = 0.7$ for M_1 and $\theta_1 = 0.2$, $\theta_2 = 0.7$, $\theta_3 = 4$ for M_2 . We simulated a few data sets for each model.

Figure 4: Showing two subunits of a complex. The two mutations are indicated in red. One possibility is that the complex cannot be formed if the two mutations are present, but still be partially functional if there is only one mutation.



3.2 Results

From the simulation described above, we get the genotypes of 400 cases and 400 controls for each SNP present in the HapMap sample in the CEPH population. For each genotype we computed its χ^2 -score and its asymptotic p-value using R (<http://www.r-project.org/>) by simulating the distribution. Indeed, since some counts are small, the actual distribution is not well approximated by a χ^2 -distribution. We then kept the SNPs with the highest score. We will briefly mention the results under model M_0 and give more details for the other two models.

Under both models M_1 and M_2 , after filtering with a marker-wise approach, we found the two causative SNPs among the three highest χ^2 -score SNPs. The third SNP, which has an unexpected high χ^2 -score, is the SNP lying only 250 bp apart from the chosen causative SNP on chromosome 13 (see Figure 2). So this result can be explained by linkage disequilibrium. Finally, it is important to note that under the additive model the SNPs were significantly correlated with the disease also after Bonferroni correction. However, this was not the case under the multiplicative model. In other words, in a real case study a marker-wise approach would have missed the causative SNPs under the multiplicative model.

Under the model with additive effects the data matrix given in Table 6 was generated. The χ^2 -statistic corresponding to this data matrix is 73.06. To test whether this value is significant or not, the algorithm described in Section 2.4 is followed. First, we computed the Markov basis for testing for no 3-way interactions in $3 \times 3 \times 2$ tables. Then three Markov chains with a length of 100'000 iterations each with different starting values were generated and the tools described in Gilks et al. (1995) were used to assess convergence of the chains. This included plotting the running mean of multiple sequences with overdispersed starting points, analyzing the Gelman-Rubin statistic and the autocorrelations. After discarding an initial burn-in of 2'000 iterations, only every tenth sample of each chain was saved due to high autocorrelations. These samples of the three chains were then combined for computing the posterior

density shown in Figure 5. So under the posterior distribution resulting from the MCMC algorithm this value is highly significant.

Under the multiplicative model we got similar results. The generated data matrix is shown in Table 7. The χ^2 -statistic corresponding to this data matrix is 72.67 and the posterior distribution of the χ^2 -statistic is shown in Figure 6. So the resulting χ^2 -statistic is highly significant.

Under model M_0 the two SNPs with the highest χ^2 -score were two random SNPs, but their p-value was not significant after Bonferroni correction. The data matrix simulated under this model is given in Table 8. The χ^2 -statistic corresponding to this data matrix is 14.83 and the posterior distribution of the χ^2 -statistic is shown in Figure 7. So this result is slightly significant.

Finally, we also performed another control study, where we used one causative SNP and another SNP chosen at random. In this case we got the data matrix shown in Table 9. The corresponding χ^2 -statistic is 0.87. Under the posterior distribution shown in Figure 8, this value results in a p-value which is not significant.

Table 6: Showing $3 \times 3 \times 2$ contingency table with the simulated data under the additive effects model.

		Non-diseased individuals		
		Marker 2:		
		0	1	2
Marker 1:	0	127	72	14
	1	99	47	11
	2	17	12	1

		Diseased individuals		
		Marker 2:		
		0	1	2
Marker 1:	0	0	102	19
	1	132	82	18
	2	26	18	3

Figure 5: Posterior density of the χ^2 -statistic under the additive effects model.

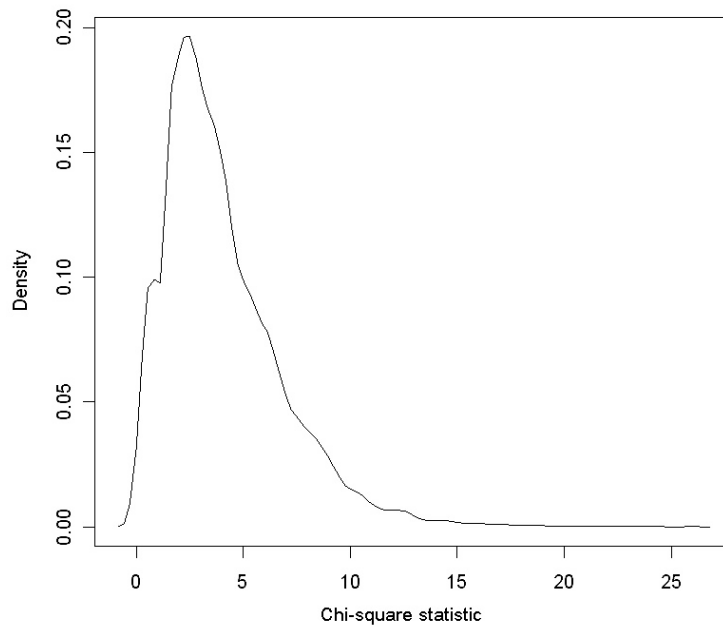


Table 7: Showing $3 \times 3 \times 2$ contingency table with the simulated data under the multiplicative effects model.

		Non-diseased individuals		
		Marker 2:		
		0	1	2
Marker 1:	0	136	56	10
	1	105	56	9
	2	18	10	0

		Diseased individuals		
		Marker 2:		
		0	1	2
Marker 1:	0	0	71	12
	1	140	101	13
	2	42	18	3

Figure 6: Posterior density of the χ^2 -statistic under the multiplicative effects model.

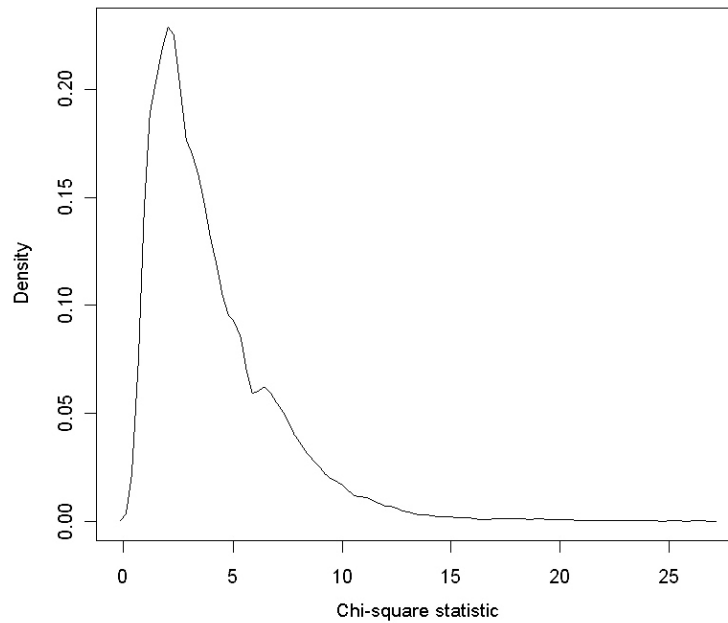


Table 8: Showing $3 \times 3 \times 2$ contingency table with the simulated data under the control model M_0 .

		Non-diseased individuals		
		Marker 2:		
		0	1	2
Marker 1:	0	130	106	17
	1	68	49	8
	2	16	3	3

		Diseased individuals		
		Marker 2:		
		0	1	2
Marker 1:	0	133	85	27
	1	69	56	7
	2	10	13	0

Figure 7: Posterior density of the χ^2 -statistic under the control model M_0 .

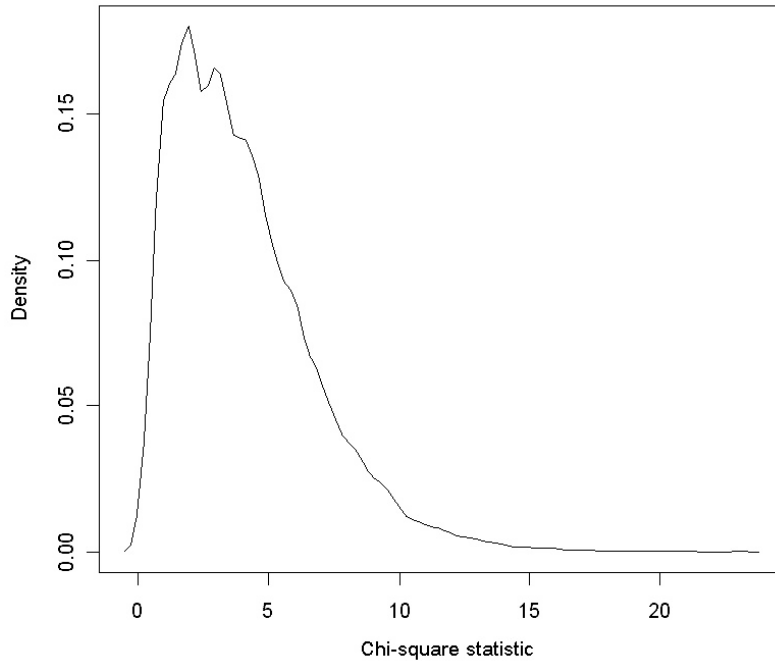
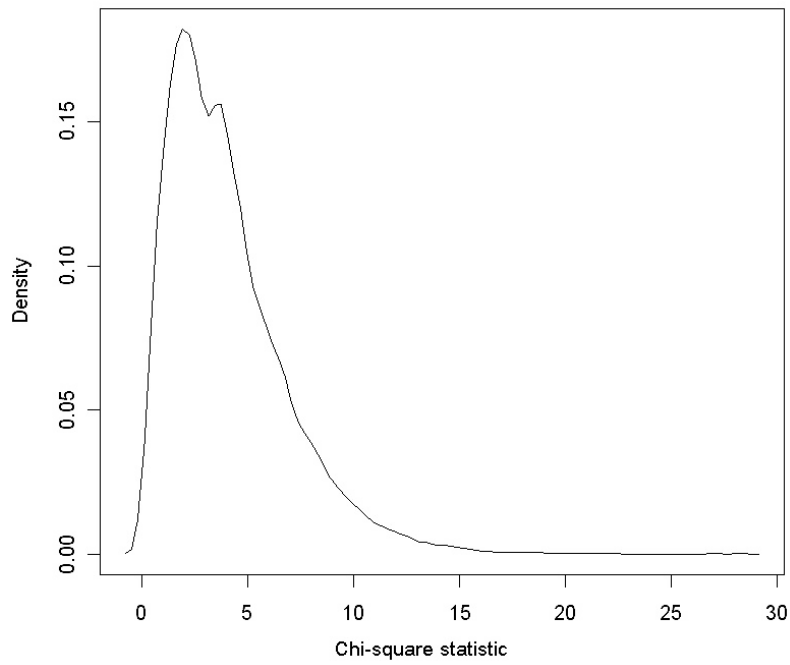


Table 9: Showing $3 \times 3 \times 2$ contingency table with the simulated data under the second control study.

		Non-diseased individuals		
		Marker 2:		
		0	1	2
Marker 1:	0	90	72	10
	1	87	77	14
	2	25	21	4

		Diseased individuals		
		Marker 2:		
		0	1	2
Marker 1:	0	36	108	24
	1	41	119	31
	2	6	27	8

Figure 8: Posterior density of the χ^2 -statistic in the second control study.



4 Discussion and Future Directions

Under both models, the model with additive effects and the model with multiplicative effects, we rejected the null hypothesis of no 3-way interactions. This means, first, that we found the two markers which are associated with the disease, whereas the marker-wise approach fails to detect the two causative SNPs. And second, that under both models an interaction between the two SNPs is present. Note that these two SNPs lie on different chromosomes. So this correlation is not due to linkage disequilibrium. Under the multiplicative model the highly significant result can be explained by epistasis. However, under the additive model the correlation found between the two markers is very likely to be caused by the non-randomness in sampling.

The first control study gives a slightly significant result, which is unexpected. In order to find out the reasons, more such control studies need to be performed. However, the second control study, gives a non-significant result.

Concluding, we think that tests for interactions between markers need to be performed in multidimensional tables. In addition, we have seen that the problem of non-random sampling can alter the results and should be taken into account in disease association studies.

Further research should include how to choose the SNPs that we further want to analyze in a multidimensional table. One approach would be to use the recently developed tool of logic regression. Another interesting question would be to study which forms of epistasis can be distinguished more easily than others and if there is a way to distinguish between additive effects and epistasis. However, this might be a hard problem due to the correlation introduced by the non-randomness in sampling individuals for disease association studies.

Moreover a thorough power analysis should be performed in order to characterize the limits of our test. Indeed, we chose a favorable setting for our simulations. This included high penetrance of the disease, a relatively high value of the interaction parameter θ_3 in the multiplicative model, only two causative loci and no environmental factors. It would also be of importance to compare our approach to the simple χ^2 test between haplotypes and the disease as well as to other existing methods (e.g. Albrechtsen et al. (2007) and citations therein).

Finally, we should try to characterize more thoroughly the possible mechanisms that can lead to correlation between markers in association studies.

Acknowledgements

We would like to thank Lior Pachter, Yun Song, Monty Slatkin and Anders Albrechtsen for helpful discussions.

References

- Albrechtsen, A. (2007). A bayesian multilocus association method: allowing for higher-order interaction in association studies. *Genetics*, 176, 1197–1208.
- Bishop, Y., Fienberg, S., Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.
- Cordell, H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11, 2463–2468.
- Diaconis, P., Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26, 363–397.
- Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (Eds) (1995). *Markov chain Monte Carlo in practice*. Chapman & Hall, London.
- Ha, H. T. (2002). Box-shaped matrices and the defining ideal of certain blowup surfaces. *Journal of Pure and Applied Algebra*, 167, 203–224.
- Marchini J., Donnelly P., Cardon L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37, 413–417.
- Sturmfels, B. (1995). *Groebner Bases and Convex Polytopes*. University Lecture Series, Volume 8, American Mathematical Society.
- The International HapMap Consortium. (2003) The International HapMap Project. *Nature* 426, 789-796.
- Wright, F. A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., Barry, W. T., Pardo-Manuel de Villena, F., Sullivan, P. F., Wilhelmsen, K. C., and Zou, F. (2007). Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics* 23, 2581–2588.