A strategy for detecting multiple trait loci in disease association studies

Anna-Sapfo Malaspinas and Caroline Uhler

December 11, 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. Our method is based on an extended Fisher's exact test for multidimensional contingency tables. To perform this test, a Markov chain is constructed on the space of multidimensional contingency tables using Markov bases. Concluding, we test our method on simulated data, showing that we can detect interacting loci where other methods fail to do so.

1 Overview

1.1 Introduction

The most common causes of mortality in developed countries are conditions such as cancer, heart disease, and diabetes, which have significant genetic components. Therefore, the mapping of genes involved in such complex diseases represents a major goal of human genetics. However, genetic variants associated with complex diseases are hard to detect because they show very little effect independently and have a low penetrance. Those genetic variants most likely interact to produce the disease phenotype. But until now there has been only little evidence for the presence of multilocus interaction in complex diseases.

Recent development of methods to screen hundreds of thousands of SNPs has allowed the discovery of over 50 disease susceptibility loci with marginal effects [7]. Genome-wide association studies have hence proved to be fruitful in understanding complex multifactorial traits. The quasi absence of reports of interacting loci however, shows the need for better methods for detecting not only marginal effects of specific loci, but also interactions of loci.

Although some progress in detecting interactions has been achieved in the last few years using simple log-linear models, those methods remain inefficient to detect interactions for large-scale data [1]. Recent studies have revealed the importance of searching for interactions between multiple loci for various diseases (e.g. [6]). Many 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 2, trying to cast them in a biological framework.

In the method described in this paper, we suggest to first reduce the potential interacting SNPs to a small number by filtering all SNPs genome-wide with a single locus approach, following what has been suggested by Marchini et al. in [6]. 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. For some models of interaction, Marchini et al. show that the two-stage approach outperforms the single-locus search and performs at least as good as when testing for k-way interactions of all subsets of k SNPs. This result suggests that a two-stage test is a reasonable approach.

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, where each SNP is tested for interaction with the disease status. Bonferroni corrections for the p-value are in general used to account for the large number of tests performed. We will use this approach as a first filter in our method to reduce the large number of SNPs to a small subset which will then be further analyzed. It is desirable to test various associations between a selection of markers by an exact test. We will use an extended version of Fisher's exact test to test for various associations within a selection of markers.

In what follows, we first discuss the term 'interaction', as it has different meanings in different contexts. We then shortly review tests used for detecting interaction.

1.2 Interaction between markers

In a biological context, interaction between markers (or SNPs) is in general used as synonym for *epistasis*. Cordell [4] gives a broad definition: "Epistasis refers to departure from 'independence' 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 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.

1.3 Tests for interaction in contingency tables

There are two widely used tests for detecting interaction within contingency tables. One is the χ^2-test. Under independence it is a well-known fact that the χ^2-statistic is asymptotically χ^2-distributed. However, note that this approximation is problematic with small counts. This is probable in contingency tables resulting from disease association studies.

The other widely used test is Fisher's exact test. As its name suggests, it has the advantage of being exact. But it is a permutation test and therefore computationally more intensive. In order to compute the p-value of a given two-dimensional table, all tables with the same row and column sums have to be found. For tables with large total counts or tables of higher dimension enumerating all possible tables with the given margins is impossible. So we want to sample these tables and approximate the exact distribution. Diaconis and Sturmfels [5] show how Markov bases can be used to perform Fisher's exact test on multidimensional tables, where the exact distribution is approximated by MCMC.

1.4 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 status D. All proofs and further explanations can be found in [3].

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 [3] 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 3-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.

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. In [3] it is shown 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 [3].

For disease association studies the no 3-way interaction model, namely the model (XY, XD, YD), is particularly interesting. We will use this model in our testing procedure presented in Section 3.

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)	$(n_{i.k})$, $(n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{i.k}n_{j.k}}{(n_{..k})}$
(XY, XD, YD)	$(n_{ij.})$, $(n_{i.k})$, $(n_{.jk})$	Iterative proportional fitting

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 testing for interaction between markers necessarily implies working with multidimensional contingency tables and cannot be performed by collapsing the multidimensional tables to two-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 $(n_{ij.})$ 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_{110}	$n_{11.}$
Total:		$n_{..0}$	$n_{..1}$	$n_{...}$

In the following, we first describe how we simulated the SNP data for cases and controls and how we selected a small subset of SNPs, in our case two SNPs, which are most correlated with the disease. In Section 3, we describe our testing procedure, which is based on the extended Fisher's exact test described in [5]. In the same section we also present a power analysis of our method. Finally, the last section is devoted to a comparison of our method to BEAM, a program for Bayesian inference of epistatic interactions in case-control studies [2].

2 Simulation of SNP data

We simulated SNP data for 400 cases and 400 controls using an existing simulator called hapsample [8]. The simulations were restricted to the SNPs typed with the Affy CHIP on chromosome 9 and chromosome 13, which include about 10000 SNPs per individual. On each of these chromosomes we selected one SNP to be causative. The SNPs were chosen to be far apart from any other marker (at least 20'000bp apart). The probability of being diseased given the genotypes at the causative loci is given as an input to the simulator.

We simulated data under four different models of interaction: a control model, an additive model, a multiplicative model with weak interaction and a multiplicative model with strong interaction. The parametrization is given in the following tables.

- **Control model:**

$\frac{\mathbb{P}(D=1\|\text{genotype})}{\mathbb{P}(D=0\|\text{genotype})}$	0	1	2
0	ϵ	ϵ	ϵ
1	ϵ	ϵ	ϵ
2	ϵ	ϵ	ϵ

- **Additive model:**

$\frac{\mathbb{P}(D=1\|\text{genotype})}{\mathbb{P}(D=0\|\text{genotype})}$	0	1	2
0	ϵ	$\epsilon\beta$	$\epsilon\beta^2$
1	$\epsilon\alpha$	$\epsilon\alpha\beta$	$\epsilon\alpha\beta^2$
2	$\epsilon\alpha^2$	$\epsilon\alpha^2\beta$	$\epsilon\alpha^2\beta^2$

- **Multiplicative model:**

$\frac{\mathbb{P}(D=1\|\text{genotype})}{\mathbb{P}(D=0\|\text{genotype})}$	0	1	2
0	ϵ	$\epsilon\beta$	$\epsilon\beta^2$
1	$\epsilon\alpha$	$\epsilon\alpha\beta\delta$	$\epsilon\alpha\beta^2\delta^2$
2	$\epsilon\alpha^2$	$\epsilon\alpha^2\beta\delta^2$	$\epsilon\alpha^2\beta^2\delta^4$

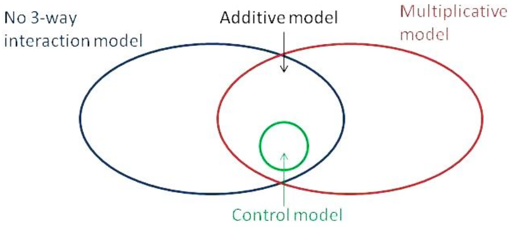
We chose the values $\epsilon = 0.05$ for the control model; $\epsilon = 0.05$, $\alpha = 2$, $\beta = 3.5$ for the additive model; $\epsilon = 0.05$, $\alpha = 2$, $\beta = 3.5$, $\delta = 3$ for the multiplicative model with weak interaction; $\epsilon = 0.05$, $\alpha = 2$, $\beta = 3.5$, $\delta = 6$ for the multiplicative model with strong interaction. We simulated a total of 200 potential association studies with 400 cases and 400 controls for each of the disease models and three different minor allele frequencies. For each study, out of the 10000 SNPs, we selected the two SNPs which are most correlated with the disease status using a χ^2 test in a standard single locus approach. This leads to a $3 \times 3 \times 2$ contingency table.

3 Hypothesis testing

As described in the previous section, we simulated data under the control model, the additive model and the multiplicative model. For performing hypothesis testing it is important to understand how the different models are nested within each other. This is described in the following figure. It is important to note that the additive model corresponds to the whole intersection of the no 3-way interaction model with the multiplicative model, and the control model is nested within the additive model.

Our goal is to find epistasis when present and get a negative result for data simulated under the control model or the additive model. As a first test we perform the no 3-way interaction test on data sets simulated under the four models presented above.

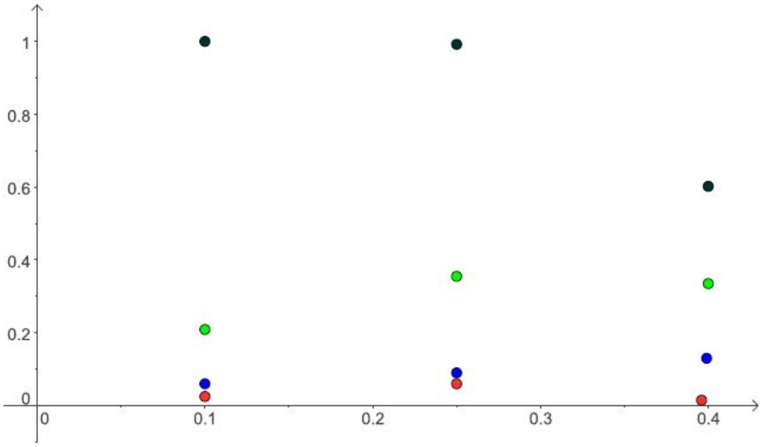
Figure 1: Nesting of the four models under consideration.



3.1 No 3-way interaction test

The null hypothesis for this test is no 3-way interaction. So we computed the Markov basis under this model and used this basis to hop between contingency tables with fixed minimal sufficient statistics and performed an MCMC analysis of the data. We computed the posterior distribution of the χ^2-statistic and approximated the exact p-value of the simulated contingency table by the resulting posterior distribution. The results are summarized in the following power analysis, where we report the percentage of how often the null hypothesis of no 3-way interaction has been rejected. For this analysis we simulated 200 contingency tables per point in the figure.

Figure 2: Power analysis of the no 3-way interaction test. The red dots correspond to simulations under the control model with $\epsilon = 0.05$, the blue dots to simulations under the additive model with $\epsilon = 0.05$, $\alpha = 2$, $\beta = 3.5$, the light green dots to simulations under the multiplicative model with weak interaction with $\epsilon = 0.05$, $\alpha = 2$, $\beta = 3.5$, $\delta = 3$, and the dark green dots to the multiplicative model with strong interaction with $\epsilon = 0.05$, $\alpha = 2$, $\beta = 3.5$, $\delta = 6$.



We see in Figure 2 that we have very few false positives under the control model and the additive model. So our test has a high specificity. Under the multiplicative model we simulated two data sets. For a low multiplicative effect, this test performs poorly and has a low sensitivity. However, we see that for a large multiplicative effect, our test detects epistasis very accurately.

If one is interested in detecting also purely additive effects, one can perform a test using the additive model as null hypothesis and the χ^2-statistic as test statistic on the contingency tables which have previously been classified as non-epistatic. In this case, the testing procedure would be as follows: First, perform the no 3-way interaction test. If this test is rejected, conclude that there is an epistatic effect. If the null hypothesis is not rejected, conclude that there is no epistatic effect. Then perform a test using the additive model as null hypothesis on the contingency tables which have previously been classified as non-epistatic.

4 Comparison to BEAM

To validate our method, we analyzed the same simulated data sets with BEAM, a program that aims to detect epistatic interactions in disease association studies. It has been shown that BEAM is significantly more powerful than existing approaches [2]. In addition, similar to the method we suggest in this paper, BEAM is able to perform genome-wide association studies with many thousands of markers.

In the following figures, we compare the performance of BEAM to our method based on an extended version of Fisher's exact test. We report the percentage of how often the null hypothesis of no epistasis has been rejected.

In Figure 3 we compare the performance of our method to BEAM under the control model. Our method makes slightly more type I errors. However, the type I error rate of our method is still below 5%.

In Figure 4 we analyze the performance of the two methods under the additive model. Also in this case the probability of rejecting the null hypothesis of no 3-way interaction is higher in our method compared to BEAM. So our method has a higher type I error rate. However, when analyzing the power of the two methods to detect the marginal effects of the two causative SNPs, we see that BEAM does not detect the two causative SNPs in any simulation. Our method does much better in this perspective. The two causative SNPs are detected in 30 to 80% of all simulations.

In Figure 5 we measure the power of the two tests to detect epistasis when present. We report, how often epistatic effects are detected, i.e. how often the null hypothesis of no 3-way interaction is rejected. We see that BEAM does not detect interaction in any of the simulations. In comparison, with our method we are able to detect interaction in 30 to 40% of the simulations. Also when comparing the power to detect marginal effects of the two causative SNPs, our method performs much better. It has a power of nearly one compared to a power of zero to 30% for BEAM.

Figure 3: Type I error of no 3-way interaction test compared to BEAM for simulations under the control model with $\epsilon = 0.05$. The red dots correspond to our method, whereas the blue dots are the results with BEAM.

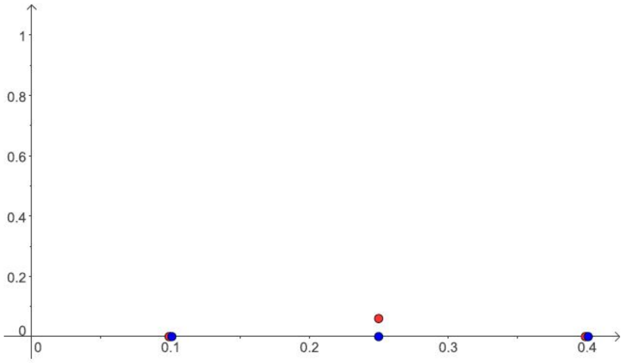


Figure 4: Type I error of no 3-way interaction test compared to BEAM for simulations under the additive model with $\epsilon = 0.05$, $\alpha = 2$, $\beta = 3.5$. The red dots correspond to our method, whereas the blue dots are the results with BEAM. The orange dots represent the probability of detecting marginal effects for both causative SNPs under our method, and the turquoise dots under BEAM.

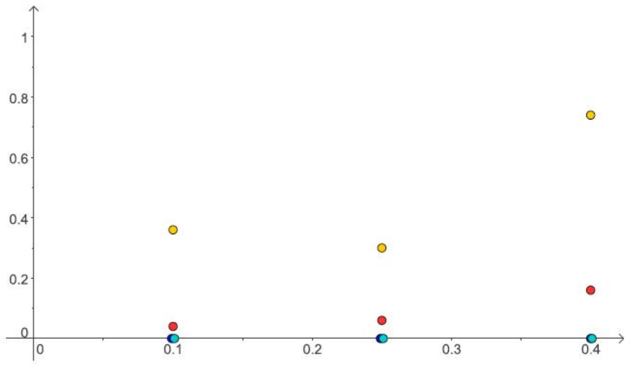
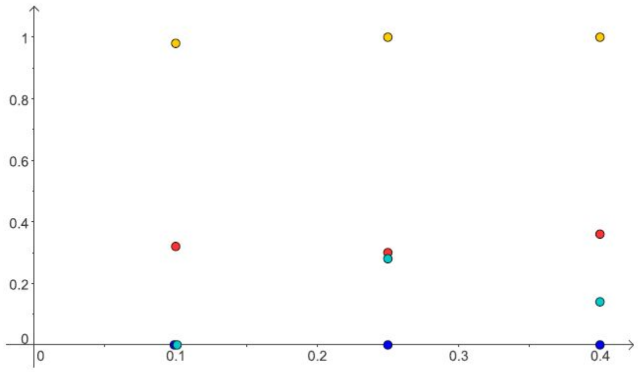


Figure 5: Power analysis of no 3-way interaction test compared to BEAM for simulations under the multiplicative model with weak interaction with $\epsilon = 0.05$, $\alpha = 2$, $\beta = 3.5$, $\delta = 3$. The red dots correspond to our method, whereas the blue dots are the results with BEAM. The orange dots represent the probability of detecting marginal effects for both causative SNPs under our method, and the turquoise dots under BEAM.



Finally, we should also mention that for this comparison we had to select only 1000 SNPs out of the 10000 SNPs simulated for the analysis in Section 3. The reason is that BEAM takes 580 minutes for the analysis of one data set with 10000 SNPs and 500 cases and controls. In comparison, for the same analysis our method takes 3 minutes.

5 Discussion

It has been argued that the χ^2 goodness-of-fit test has little power to detect marginal effects when there is high interaction between causative SNPs (e.g. [2]). For the above analysis we nevertheless used the χ^2-test as filter. We asked ourselves if it might be useful to exchange our filter with a more sophisticated method or if it would suffice to choose say the 10 SNPs which are most correlated with the disease by the χ^2-test and then perform the no 3-way interaction test on all pairs. We were debating to use BEAM as a filter instead of the χ^2-test. We evaluated both methods mimicking a real complex disease situation.

In order to mimic a real complex disease situation we computed the effect size, λ, and the prevalence of the disease, π, for each of our models. The effect size for a particular locus i is defined as follows:

$$\lambda_i = \frac{p(D=1|g_i=1)}{p(D=0|g_i=1)}\frac{p(D=0|g_i=0)}{p(D=1|g_i=0)} - 1,$$

and the prevalence as follows:

$$\pi = \sum_{g_1,g_2} p(D|g_1,g_2)p(g_1,g_2),$$

where g_1 and g_2 are the genotypes at each of the causative loci. For common diseases the effective size is known to be around 0.3 and the prevalence is generally 5%. We realized that the parameters chosen for the simulation study in Section 3 and 4 lead to values of λ and π, which were higher then the respective values for common diseases. So for each of the models considered, we then chose parameters α, β, δ, and ϵ resulting in $\lambda = 0.3$ and $\pi = 0.05$. We ran more simulations with those new parameters. However, our filter, the χ^2-test, had little power to recover the causative SNPs. Interestingly, BEAM did not perform better in this task. In addition, as mentioned earlier, the long running time of BEAM makes it unfeasible to use BEAM as filter on the whole genome. So we have to conclude that the simple χ^2-test is better suited for our method then using a more complex method such as BEAM.

So we can conclude that the method presented in this paper is more powerful than other existing methods. However, as a final step in our project it would be desirable to find filters, which are able to detect the marginal effects of causative SNPs for simulations that match real life situations better.

References

[1] Albrechtsen, A. (2007). A bayesian multilocus association method: allowing for higher-order interaction in association studies. *Genetics*, 176, 1197–1208.

[2] Zhang, Y., Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39, 1167–1173.

[3] Bishop, Y., Fienberg, S., Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.

[4] 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.

[5] Diaconis, P., Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26, 363–397.

[6] Marchini J., Donnelly P., Cardon L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37, 413–417.

[7] McCarthy M. I., Abecasis G. R., Cardon L. R., Goldstein D. B., Little J., Ioannidis J. P. A., Hirschhorn J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9, 356–369.

[8] 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.