JOURNAL OF ALGEBRAIC STATISTICS
Vol. 2, No. 1, 2011, 36-53
ISSN 1309-3452 – www.jalgstat.com

Detecting epistasis via Markov bases

Anna-Sapfo Malaspinas [1,*], Caroline Uhler [2,†,*]

[1] *Department of Integrative Biology, University of California at Berkeley*
[2] *Department of Statistics, University of California at Berkeley*

Abstract. Rapid research progress in genotyping techniques have allowed large genome-wide association studies. Existing methods often focus on determining associations between single loci and a specific phenotype. However, a particular phenotype is usually the result of complex relationships between multiple loci and the environment. In this paper, we describe a two-stage method for detecting epistasis by combining the traditionally used single-locus search with a search for multiway interactions. Our method is based on an extended version of Fisher's exact test. To perform this test, a Markov chain is constructed on the space of multidimensional contingency tables using the elements of a Markov basis as moves. We test our method on simulated data and compare it to a two-stage logistic regression method and to a fully Bayesian method, showing that we are able to detect the interacting loci when other methods fail to do so. Finally, we apply our method to a genome-wide data set consisting of 685 dogs and identify epistasis associated with canine hair length for four pairs of single nucleotide polymorphisms (SNPs).

2000 Mathematics Subject Classifications: 62P10, 62F03, 92B05

Key Words and Phrases: Epistasis, Markov basis, association studies, sparse contingency tables, Fisher's exact test

1. Introduction

Conditions with genetic components such as cancer, heart disease, and diabetes, are the most common causes of mortality in developed countries. 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. Indeed, only a small portion of the heritability of complex diseases can be explained by the variants identified so far. This led to several hypotheses (see e.g. [19]). One of them is that most common diseases are caused by several rare variants with low effects, rather than a few common variants with large effects ([24]). Another hypothesis is that the variants interact in order to produce the disease phenotype and independently only explain a small fraction

[*] Anna-Sapfo Malaspinas is supported by a Janggen-Poehn Fellowship.
[†] Caroline Uhler is supported by an International Fulbright Science and Technology Fellowship.
[*] Corresponding author.

Email addresses: sapfo@berkeley.edu (A. S. Malaspinas), cuhler@stat.berkeley.edu (C. Uhler)

of the genetic variance. In this paper, we mainly focus on the interaction hypothesis, but we will also discuss the relevance of our method to the rare variant hypothesis along the way.

Recent development of methods to screen hundreds of thousands of single nucleotide polymorphisms (SNPs) has allowed the discovery of over 50 disease susceptibility loci with marginal effects ([23]). Genome-wide association studies have hence proven 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, these methods remain inefficient to detect interactions for large-scale data ([2]).

Various models of interaction have been presented in the past, for example the additive model and the multiplicative model. The former model assumes that the SNPs act independently, and a single marker approach seems to perform well. In the multiplicative model, 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 SNP. We will discuss such models in more detail in Section 2.1. A complete classification of two-locus interaction models has been given in [15].

In the method described in this paper, we first reduce the potential interacting SNPs to a small number by filtering all SNPs genome-wide with a single locus approach. The loci achieving some threshold are then further examined for interactions. Such a two-stage approach has been suggested and implemented in [20]. For some models of interaction, they show that the two-stage approach outperforms the single-locus search and performs at least as well as when testing for interaction within all subsets of k SNPs.

Single locus methods consider each SNP individually and test for association based on differences in genotypic frequencies between case and control individuals. Widely used methods for the single-locus search are the χ^2 goodness-of-fit test or Fisher's exact test together with a Bonferroni correction of the p-values to account for the large number of tests performed. We suggest using Fisher's exact test as a first stage to rank the SNPs by their p-value and select a subset of SNPs, which is then further analyzed. Under the rare variant hypothesis the resulting contingency tables are sparse and it is desirable to test for interactions within the selected subset using an exact test. We suggest using Markov bases for this purpose. A Markov basis is a set of Markov chain moves, which connects a contingency table to all tables with the same margins (see e.g. [7]).

In Section 2, we define three models of interaction and present our algorithm for detecting epistasis using Markov bases in hypothesis testing. In Section 3, we test our method on simulated data and make a comparison to logistic regression and BEAM, a Bayesian approach ([31]). Finally, we run our algorithm on a genome-wide dataset from dogs ([5]) to test for epistasis related to canine hair length.

2. Method

2.1. Models of interaction

In this paper, we mainly study the interaction between two SNPs and a binary phenotype, as for example the disease status of an individual. However, our method can be easily generalized for studying interaction between three or more SNPs and a phenotype with three or more states. We show a generalization in Section 3.4, where we analyze a genome-wide dataset from dogs and, inter alia, test for interaction between three SNPs and a binary hair length phenotype (short hair versus long hair).

The binary phenotype is denoted by D, taking values 0 and 1. We assume that the SNPs are polymorphic with only two possible nucleotides. The two SNPs are denoted by X and Y, each with genotypes taking values 0, 1 and 2 representing the number of minor alleles. We investigate three different models of interaction: a control model, an additive model, and a multiplicative model. The parameterization is given in the following tables showing the odds of having a specific phenotype

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

- **Control model:**

		Y		
		0	1	2
	0	ϵ	ϵ	ϵ
X	1	ϵ	ϵ	ϵ
	2	ϵ	ϵ	ϵ

- **Additive model:**

		Y		
		0	1	2
	0	ϵ	$\epsilon\beta$	$\epsilon\beta^2$
X	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:**

		Y		
		0	1	2
	0	ϵ	$\epsilon\beta$	$\epsilon\beta^2$
X	1	$\epsilon\alpha$	$\epsilon\alpha\beta\delta$	$\epsilon\alpha\beta^2\delta^2$
	2	$\epsilon\alpha^2$	$\epsilon\alpha^2\beta\delta^2$	$\epsilon\alpha^2\beta^2\delta^4$

These three models can also be expressed as log-linear models. We denote the state of X by i, the state of Y by j, and the state of D by k. If n_{ijk} describes the expected

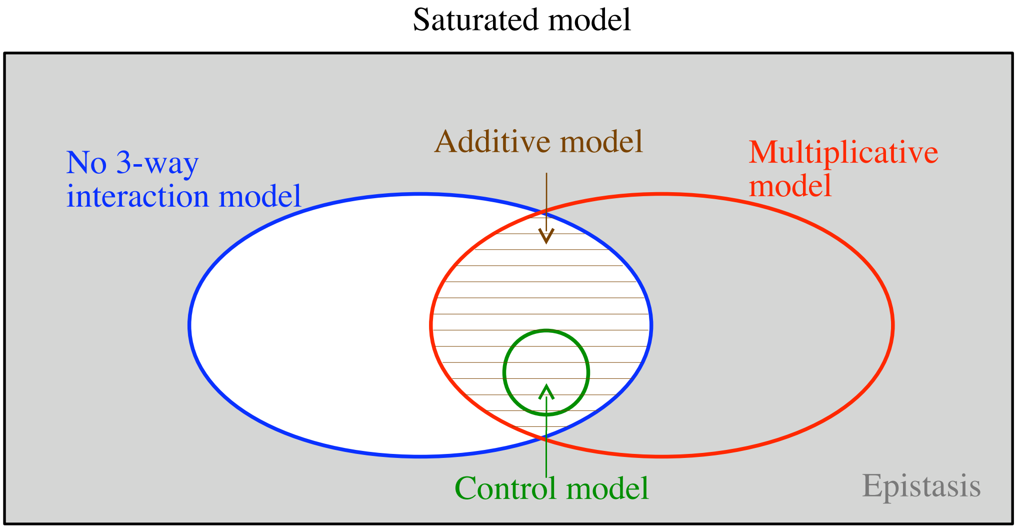

Figure 1: Nesting relationship of the control model, the additive model, and the multiplicative model. The intersection of the no 3-way interaction model with the multiplicative model corresponds to the additive model. The shading indicates the presence of epistasis.

cell counts in a 3 × 3 × 2 contingency table, then the three models can be expressed in the following way, where the γ terms represent the effects the variables have on the cell counts (e.g. γ_i^X represents the main effect for X), and α, β, δ, and ϵ are defined by the odds of having a specific phenotype shown in the above tables:

Control model: $\log(n_{ijk}) = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_{ij}^{XY} + k\log(\epsilon)$

Additive model: $\log(n_{ijk}) = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_{ij}^{XY} + k\log(\epsilon) + ik\log\alpha + jk\log\beta$

Multiplicative model: $\log(n_{ijk}) = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_{ij}^{XY} + k\log(\epsilon) + ik\log\alpha + jk\log\beta + ijk\log\delta$

Note that in the additive model the interaction effect for SNP X (SNP Y) and the disease status is additive with respect to the number of causative SNPs i (j), whereas in the multiplicative model there is an additional 3-way interaction effect between SNPs X, Y, and the disease status, which is multiplicative in the number of causative SNPs i, j. From the representation as log-linear models we can deduce the nesting relationship shown on the Venn diagram in Figure 1. Note that the additive model corresponds to the intersection of the no 3-way interaction model ($\log(n_{ijk}) = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^D + \gamma_{ij}^{XY} + \gamma_{ik}^{XD} + \gamma_{jk}^{YD}$) with the multiplicative model, and the control model is nested within the additive model.

In a biological context, interaction between markers (or SNPs) is usually used as a synonym for *epistasis*. Cordell [6] 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 (i.e. $\log(\delta) \neq 0$ in the multiplicative model).

In contrast, in a mathematical context interaction is used as synonym for *dependence*. Two markers are said to be interacting if they are dependent, 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 associated with a specific phenotype. In what follows, we will use the term interaction as synonym for dependence and the term epistasis with respect to a specific phenotype synonymously to the presence of a k-way interaction ($k \geq 3$) between $k-1$ SNPs and a specific phenotype. The epistatic models are indicated by the shading in Figure 1.

2.2. Algorithm

The χ^2 goodness-of-fit-test is the most widely used test for detecting interaction within contingency tables. Under independence the χ^2 statistic is asymptotically χ^2 distributed. However, this approximation is problematic when some cell counts are small, which is often the case in contingency tables resulting from association studies and particularly problematic under the rare variant hypothesis. 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. For tables with large total counts or tables of higher dimension, enumerating all possible tables with given margins is not feasible.

Diaconis and Sturmfels [7] describe an extended version of Fisher's exact test using Markov bases. A Markov basis for testing a specific interaction model is a set of moves connecting all contingency tables with the same sufficient statistics. A Markov basis allows the construction of a Markov chain on the set of contingency tables with given margins and computing the p-value of a given contingency table using its stationary distribution. Such a test can be used for analyzing multidimensional tables with large total counts. In addition, it has been shown in [7] that the stationary distribution is a good approximation of the exact distribution of the χ^2-statistic even for very sparse contingency tables, leading to a substantially more accurate interaction test than the χ^2-test for sparse tables. Useful properties of Markov bases can be found in [9].

A Markov basis of the null model can be computed using the software `4ti2` ([1]). All Markov bases mentioned in this paper can be found on our website[§]. Then a Markov chain is started at the observed $3 \times 3 \times 2$ data table and elements of the Markov basis are used as moves in the Metropolis-Hastings steps. At each step the χ^2 statistic is computed. Its stationary distribution is an approximation of the exact distribution of the χ^2 statistic.

2.2.1. Interaction tests with the extended version of Fisher's exact test

In this subsection we present various hypotheses that can easily be tested using Markov bases and discuss a hypothesis that is particularly interesting for association studies. For simplicity we constrain this discussion to the case of two SNPs and a binary phenotype.

[§] http://www.carolineuhler.com/epistasis.htm

Table 1: Standard 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

Table 1 consists of the standard log-linear models on three variables. Their fit to a given data table can be computed using the extended version of Fisher's exact test. We use the notation presented in [4] to denote the different models. Interaction is assumed between the variables not separated by commas in the model. So the model (X, Y, D) in Table 1 represents the independence model, the model (XY, XD, YD) the no 3-way interaction model and the other models are intermediate models. For association studies the no 3-way interaction model (XY, XD, YD) is particularly interesting and will be used as null model in our testing procedure.

Performing the extended version of 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. If a loop is present in the model configuration as for example in the no 3-way interaction model (this model can be rewritten as (XY, YD, DX)), then there is no closed-form estimator for the cell counts (see [4]). But in this case, estimates can be achieved by iterative proportional fitting (i.e. [11]).

It is important to note that testing for epistasis necessarily implies working with multi-dimensional contingency tables and is not possible in the collapsed two-dimensional table shown above. In this table, the two SNPs are treated like a single variable and we consider their haplotype. 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 using (XY, D) as null model. In this case, the null hypothesis would be rejected even in the presence of marginal effects only, showing that testing for epistasis in Table 2 is impossible.

2.2.2. Hypothesis testing with the extended version of Fisher's exact test

Our goal is to detect epistasis when present. According to the definition of epistasis in Section 2.1 and as shown in Figure 1, epistasis is present with regard to two SNPs and a specific phenotype, when a 3-way interaction is found. So we suggest using as null

Table 2: Testing for association between haplotypes and phenotype.

		Phenotype status:		**Total:**
		0	1	
Haplotype:	00	n_{000}	n_{001}	$n_{00.}$
	01	n_{010}	n_{011}	$n_{01.}$
	10	n_{100}	n_{101}	$n_{10.}$
	11	n_{110}	n_{111}	$n_{11.}$
Total:		$n_{..0}$	$n_{..1}$	$n_{...}$

hypothesis the no 3-way interaction model and testing this hypothesis with the extended version of Fisher's exact test. The corresponding minimal Markov basis consists of 15 moves. It can be used to approximate the exact distribution of the χ^2 statistic and compute the p-value of the data table. If the p-value is lower than some threshold, we reject the null hypothesis of no epistasis.

Although in this paper we focus merely on epistasis, it is worth noting that one can easily build tests for different types of interaction using Markov bases. If one is interested in detecting whether the epistatic effect is of multiplicative nature, one can perform the extended version of Fisher's exact test on the contingency tables, which have been classified as epistatic, using the multiplicative model as null hypothesis. In this case, the corresponding minimal Markov basis consists of 49 moves. Similarly, if one is interested in detecting additive effects, one can use the additive model as null hypothesis and test the contingency tables, which have been classified as non-epistatic. In this case, the corresponding minimal Markov basis consists of 156 moves. Markov bases for all mentioned tests can be found on our website[¶]. A strength of the Markov basis approach is that each Markov basis only needs to be computed once and can then be reused.

3. Results

In this section, we first conduct a simulation study to evaluate the performance of the suggested method. We then compare our method to a two-stage logistic regression approach and to BEAM ([31]). Logistic regression is a widely used method for detecting epistasis within a selection of SNPs. BEAM is a purely Bayesian method for detecting epistatic interactions on a genome-wide scale. We end this section by applying our method to a genome-wide data set consisting of 685 dogs with the goal of finding epistasis associated with canine hair length.

3.1. Simulation study

We simulated a total of 50 potential association studies with 400 cases and 400 controls for three different minor allele frequencies of the causative SNPs and the three models of interaction presented in Section 2.1. We chose as minor allele frequencies (MAF) 0.1, 0.25

[¶] http://www.carolineuhler.com/epistasis.htm and 0.4. The parameters for the three models of interaction were determined numerically by fixing the marginal effect measured by the effect size

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

and the prevalence

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

Since only very few cases of interacting loci have been reported, little is known about the true values of the parameters in the interaction models. For our simulations, we used an effect size of $\lambda_1 = \lambda_2 = 1$ and a sample prevalence of $\pi = 0.5$. The sample prevalence corresponds to the prevalence in most disease mapping studies, where the number of cases is usually equal to the number of controls being examined. There is a trade-off between effect size and number of cases and controls needed to achieve a certain power. We chose an effect size of 1 and 400 cases and controls, similar to the two-stage approach used in [20]. Choosing in addition $\alpha = \beta$ in the additive model, and $\alpha = \beta$ and $\delta = 3\alpha$ in the multiplicative model determines all parameters of the interaction models and one can solve for α, β, δ and ϵ numerically.

The simulations were performed using HAP-SAMPLE ([30]) and were restricted to the SNPs typed with the Affy CHIP on chromosome 9 and chromosome 13 of the Phase I/II HapMap data[‖], resulting in about 10,000 SNPs per individual. On each of the two chromosomes we selected one SNP to be causative. The causative SNPs were chosen consistent with the minor allele frequencies and far apart from any other marker (at least 20,000bp apart). Note that HAP-SAMPLE generates the cases and controls by resampling from HapMap. This means that the simulated data show linkage disequilibrium and allele frequencies similar to real data.

As suggested in [20], we took a two-stage approach for finding interacting SNPs. In the first step, we ranked all SNPs according to their p-value in Fisher's exact test on the 2 × 3 genotype table and selected the ten SNPs with the lowest marginal p-values. Figure 2 shows a boxplot of the p-values of the causative SNPs for the three models under consideration and each of the three minor allele frequencies. Within the subset of the ten lowest ranked SNPs, we then tested for interaction using the extended version of Fisher's exact test with the no 3-way interaction model as null hypothesis. We generated three Markov chains with 40,000 iterations each and different starting values, and used the tools described in [12] to assess convergence of the chains. This included analyzing the Gelman-Rubin statistic and the autocorrelations. After discarding an initial burn-in of 10,000 iterations, we combined the remaining samples of the three chains to generate the posterior distribution of the χ^2 statistic.

In Figure 3 (left), we report the rejection rate of the no 3-way interaction hypothesis for each of the three minor allele frequencies. Per point in the figure we simulated 50 potential association studies. The power of our two-stage testing procedure corresponds

[‖] http://hapmap.ncbi.nlm.nih.gov/

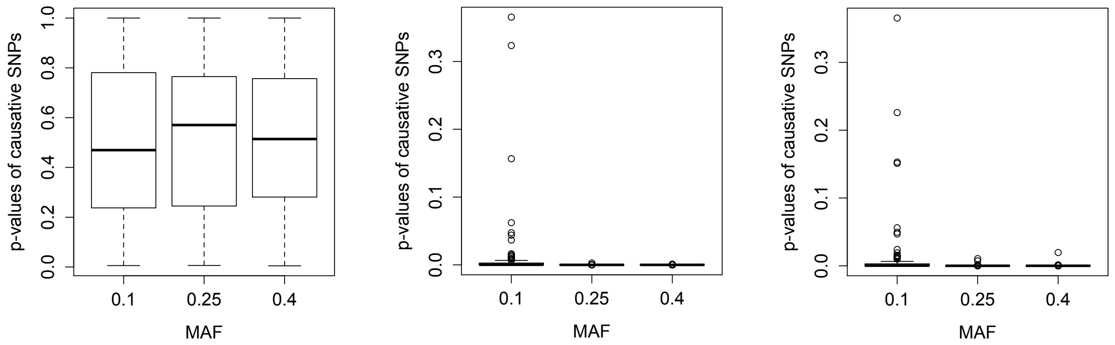

Figure 2: Boxplot of the p-values of the causative SNPs for the three minor allele frequencies and the three models of interaction, i.e. the control model (left), the additive model (middle), and the multiplicative model (right).

to the curve under the multiplicative model. The higher the minor allele frequency, the more accurately we can detect epistasis. Under the additive model and the control model, no epistasis is present. We never rejected the null hypothesis under the control model and only once under the additive model, resulting in a high specificity of the testing procedure.

We also analyzed the performance of each step separately. Figure 3 (middle) shows the performance of the first step and reports the proportion of 50 association studies, in which the two causative SNPs were ranked among the ten SNPs with the lowest p-values. Because Fisher's exact test measures marginal association, the curves under the additive model and the multiplicative model are similar.

Figure 3 (right) shows the performance of the second step in our method and reports the proportion of 50 association studies, in which the null hypothesis of no 3-way interaction was rejected using only the extended version of Fisher's exact test on the 50 causative SNP pairs.

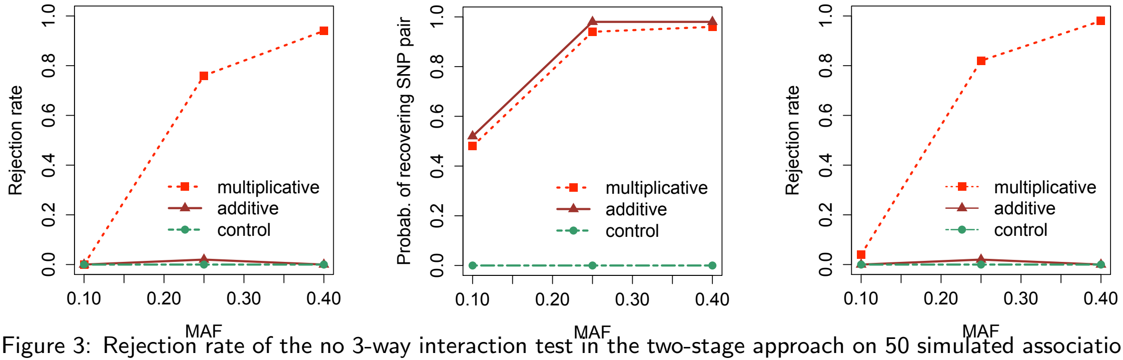

Figure 3: Rejection rate of the no 3-way interaction test in the two-stage approach on 50 simulated association studies for MAF=0.1, MAF=0.25, and MAF=0.4 (left). Proportion of 50 association studies, in which the two causative SNPs were ranked among the ten SNPs with the lowest p-values by Fisher's exact test (middle). Rejection rate of the no 3-way interaction hypothesis using only the extended version of Fisher's exact test on the 50 causative SNP pairs (right).

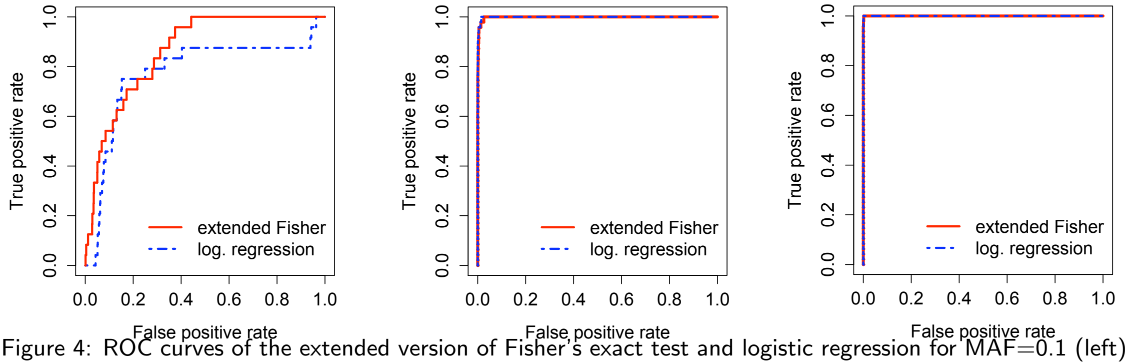

Figure 4: ROC curves of the extended version of Fisher's exact test and logistic regression for MAF=0.1 (left), MAF=0.25 (middle), and MAF=0.4 (right) based on the ten filtered SNPs.

3.2. Comparison to logistic regression

For validation, we compared the performance of our method to logistic regression via ROC curves. Logistic regression is probably the most widely used method for detecting epistasis within a selection of SNPs nowadays. We based the comparison on the simulated association studies presented in the previous section using only the simulations under the multiplicative model. The structure of interaction within this model should favor logistic regression as logistic regression tests for exactly this kind of interaction.

As before, for each minor allele frequency and each of the 50 simulation studies we first filtered all SNPs with Fisher's exact test and chose the ten SNPs with the lowest p-values for further analysis. Both causative SNPs were within the ten filtered SNPs for 19 (46) [45] out of the 50 simulation studies for MAF=0.1 (MAF=0.25) [MAF=0.4]. We then ran the extended version of Fisher's exact test and logistic regression on all possible pairs of SNPs in the subsets consisting of the ten filtered SNPs. This resulted in $50 \cdot \binom{10}{2}$ tests per minor allele frequency with 19 (46) [45] true positives for MAF=0.1 (MAF=0.25) [MAF=0.4].

Because both methods, logistic regression and our method, require a filtering step, we compared the methods only based on the ten filtered SNPs. The ROC curves comparing the second stage of our method to logistic regression are plotted in Figure 4 showing that our method performs substantially better than logistic regression for MAF=0.1 with an area under the ROC curve of 0.861 compared to 0.773 for logistic regression. For MAF=0.25 and MAF=0.4 both methods have nearly perfect ROC curves with areas 0.9986 [0.99994] for our method compared to 0.9993 [0.99997] for logistic regression for MAF=0.25 [MAF=0.4].

3.3. Comparison to BEAM

We also compared our method to BEAM, a Bayesian approach for detecting epistatic interactions in association studies ([31]). We chose BEAM, because the authors show it is more powerful than a variety of other approaches including the stepwise logistic regression approach, and it is one of the few recent methods that can handle genome-wide data.

In this method, all SNPs are divided into three groups, namely, SNPs that are not

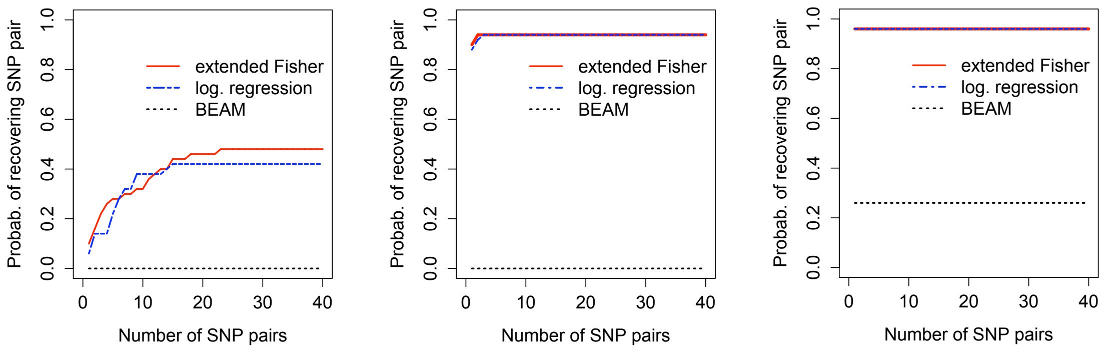

Figure 5: Proportion of simulation studies for which the interacting SNP pair belongs to the x SNP pairs with the lowest p-values for MAF=0.1 (left), MAF=0.25 (middle), and MAF=0.4 (right).

associated with the disease, SNPs that contribute to the disease risk only through main effects, and SNPs that interact to cause the disease. BEAM outputs the posterior probabilities for each SNP to belong to these three groups. The authors of [31] propose to use the results in a frequentist hypothesis-testing framework calculating the so called B-statistic and testing for association between each SNP or set of SNPs and the disease phenotype. BEAM was designed to increase the power to detect any association with the disease, and not to separate main effects from epistasis. Therefore, BEAM outputs SNPs that interact marginally **or** through a k-way interaction with the disease. This does not match our definition of epistasis since the presence of marginal effects only already gives rise to a significant result using BEAM.

We compared our method to BEAM using the B-statistic. The B-statistic is the Bayes factor which compares the alternative (where the genotypes in cases and controls follow different distributions) versus the null model (where the genotypes in cases and controls follow a common distribution). BEAM reports this statistic only for the pairs of SNPs which have a non-zero posterior probability of belonging to the third group. In addition, the B-statistic is automatically set to zero for the SNP pairs where any of the two SNPs are found to be interacting marginally with the disease. We forced BEAM to include the marginal effects into the B-statistic by choosing a significance level of zero for marginal effects. This should favor BEAM in terms of sensitivity.

We ran BEAM with the default parameters on our simulated datasets for the multiplicative model. BEAM has a long running time. It takes about 10.6 hours for the analysis of one dataset with 10,000 SNPs and 400 cases and 400 controls, whereas the same analysis takes about 0.7 hours using our method on an Intel Core 2.2 GHz laptop with 2 Gb memory. Therefore, we ran BEAM with the default parameters only on 1,000 SNPs out of the 10,000 SNPs simulated for the analysis in Section 3.1.

In contrast to BEAM, our method is a stepwise approach, which makes a comparison via ROC curves difficult. We therefore compared the performance of all three tests by plotting for a fixed number x of SNP pairs the proportion of simulation studies for which the interacting SNP pair belongs to the x SNP pairs with the lowest p-values. The resulting curves are shown in Figure 5. Although the marginal effects were not extracted, BEAM has a very high false negative rate, attributing a p-value of 1 to the majority of SNPs, interacting and not interacting SNPs.

3.4. Genome-wide association study of hair length in dogs

We demonstrate the potential of our Markov basis method in genome-wide association studies by analyzing a hair length dataset consisting of 685 dogs from 65 breeds and containing 40, 842 SNPs ([5]).

The individuals in [5] were divided into two groups for the hair length phenotype: 319 dogs from 31 breeds with long hair as cases and 364 from 34 breeds with short hair as controls. In the original study, it is shown that the long versus short hair phenotype is associated with a mutation (Cys95Phe) that changes exon one in the *fibroblast growth factor-5* (*FGF5* gene). Indeed, the SNP with the lowest p-value using Fisher's exact test is located on chromosome 32 at position 7, 100, 913 for the Canmap dataset, i.e. about 300Kb apart from *FGF5*.

We ranked the 40, 842 SNPs by their p-value using Fisher's exact test and selected the 20 lowest ranked SNPs (about 0.05%) to test for 3-way interaction. Note that all 20 SNPs are significantly correlated (p-value < 0.05) with the phenotype. We found a significant p-value (< 0.05) for four out of the $\binom{20}{2}$ pairs. These pairs together with their p-values are listed in Table 3.

The pairs include six distinct SNPs located on five different chromosomes and the two SNPs lying on the same chromosome are not significantly interacting (p-value of 0.54). This means that a false positive correlation due to hitchhiking effects can likely be avoided. Hitchhiking effects are known to extend across long stretches of chromosomes in particular in domesticated species ([26, 27, 21]) consistent with the prediction of [25].

In order to identify potential pathways we first considered genes, which are close to the six SNPs we identified as interacting. To do so, we used the dog genome available through the ncbi website[**]. Most of the genes we report here have been annotated automatically. Our strategy was to consider the gene containing the candidate SNP (if any) and the immediate left and right neighboring gene, resulting in a total of two or three genes per SNP.

Among the six significantly interacting SNPs, four are located close to genes that have been shown to be linked to hair growth in other organisms. This is not surprising, since these SNPs also have a significant marginal association with hair growth. We here report the function of these candidate genes. The two other SNPs are located close to genes that we were not able to identify as functionally related to hair growth.

First, the SNP chr30.18465869 is located close to (about 80Kb) *fibroblast growth factor 7* (*FGF7* also called *keratinocyte growth factor*, *KGF*), i.e. it belongs to the same family as the gene reported in the original study (but on a different chromosome). The FGF family members are involved in a variety of biological processes including hair development reported in human, mouse, rat and chicken (GO:0031069, [3]).

[**] http://www.ncbi.nlm.nih.gov/genome/guide/dog/, build 2.1

Table 3: Pairs of SNPs, which significantly interact with the hair length phenotype for the Canmap dataset. Question marks indicate that we were not able to identify a close-by gene which is functionally related to hair growth.

chromosome and location of SNPs	p-value	potential relevant genes
chr30.18465869, chr26.6171079	0	*FGF7*-?
chr15.44092912, chr23.49871523	0	*IGF1*-*P2RY1*
chr24.26359293, chr15.43667654	2e-04	*ASIP*-?
chr15.43667654, chr23.49871523	1e-04	?-*P2RY1*

Secondly, chr15.44092912 is located between two genes, and about 200Kb from the *insulin-like growth factor 1* gene (*IGF1*). *IGF1* has been reported to be associated with the hair growth cycle and the differentiation of the hair shaft in mice ([28]).

Thirdly, chr23.49871523 is located about 430Kb from the *purinergic receptor P2Y1* (*P2RY1*). The purinergic receptors have been shown to be part of a signaling system for proliferation and differentiation in human anagen hair follicles ([14]).

Finally, chr24.26359293 is located inside the agouti-signaling protein (gene *ASIP*), a gene known to affect coat color in dogs and other mammals. The link to hair growth is not obvious but this gene is expressed during four to seven days of hair growth in mice ([29]).

According to our analysis, *IGF1* and *P2RY1* are significantly interacting. All other pairs of interacting SNPs involve at least one SNP for which we were not able to identify a close-by candidate gene related to hair growth (see Table 3). *IGF1* has a tyrosine kinase receptor and *P2RY1* is a G-protein coupled receptor. One possibility is that these receptors cross-talk as has been shown previously for these types of receptors in order to control mitogenic signals ([8]). However, a functional assay would be necessary to establish that any of the statistical interactions we found are also biologically meaningful.

We also considered all triplets of SNPs among the 20 preselected SNPs and tested for 4-way interaction. However, we did not find any evidence for interaction among the $\binom{20}{3}$ triplets.

4. Discussion

In this paper, we proposed a Markov basis approach for detecting epistasis in genome-wide association studies. The use of different Markov bases allows to easily test for different types of interaction and epistasis involving two or more SNPs. These Markov bases need to be computed only once and can be downloaded from our website[††] for the tests presented in this paper.

The use of an exact test is of particular relevance for disease mapping studies where the contingency tables are often sparse. One example where there has been also functional validation, is a deletion associated with Crohn's disease [22]: This deletion was found to have a population frequency of 0.07, and a frequency of 0.11 in the cases [19]. So within

[††] http://www.carolineuhler.com/epistasis.htm

400 controls and under Hardy-Weinberg equilibrium, we would expect only 2 individuals to be homozygote for this deletion. This shows that also for a moderate number of cases and controls the resulting tables for disease association studies are likely to be sparse. The sparsity is even more pronounced for rare variants, defined as variants with a MAF smaller than 0.005. Current genome wide association studies are still missing these rare variants, but advances in sequencing technologies should allow to sequence these variants and appropriate statistical methods will then be necessary.

We tested our method in simulation studies and showed that it outperforms a stepwise logistic regression approach and BEAM for the multiplicative interaction model. Logistic regression has the advantage of a very short running time (3 seconds compared to 39 minutes using our method for the analysis of one dataset with 10,000 SNPs and 400 cases and controls not including the filtering step, which takes about 1 minute for both methods on an Intel Core 2.2 GHz laptop with 2 Gb memory). However, especially for a minor allele frequency of 0.1, logistic regression performs worse than our method, even when simulating epistasis under a multiplicative model, which should favor logistic regression. This difference arises because our method approximates the exact p-value well for all sample sizes while the performance of logistic regression increases with larger sample size. 400 cases and 400 controls are not sufficient to get a good performance using logistic regression for a minor allele frequency of 0.1 and it is expected to do even worse for rare variants. Another advantage of our method compared to logistic regression is that it is not geared towards testing for multiplicative interaction only, but should be able to detect epistasis regardless of the interaction model chosen. It would be interesting to compare these two methods on data sets generated by other interaction models.

BEAM on the other hand, has the advantage of not needing to filter the large number of SNPs first. However, it runs about 15 times slower than our method for our simulations and has a very high false negative rate. The difference between our results and what the authors of BEAM have found might be due to linkage disequilibrium in our data. BEAM handles linkage disequilibrium with a first order Markov chain, which will be improved in future versions (Yu Zhang, personal communication). But as of today, we conclude that this method is impractical for whole genome association studies, since linkage disequilibrium is present in most real datasets.

The limitation of our method is the need for a filtering step to reduce the number of SNPs to a small subset. Especially if the marginal association of the interacting SNPs with the disease is small, these SNPs might not be caught by the filter. However, in our simulations using Fisher's exact test as a filter seems to perform well. Another possibility is to incorporate biological information such as existing pathways ([10]) to choose a subset of possibly interacting SNPs.

We demonstrated the potential of the proposed two-stage method in genome-wide association studies by analyzing a hair length dataset consisting of 685 dogs and containing $40,842$ SNPs using the extended version of Fisher's exact test. In this dataset, we found a significant epistatic effect for four SNP pairs. These SNPs lie on different chromosomes, reducing the risk of a false positive correlation due to linkage effects. The dataset includes dogs from 65 distinct breeds. Although linkage disequilibrium has been shown to extend over several megabases within breeds, linkage disquelibrium extends only over tens of kilobases between breeds and drops faster than in human populations ([26], [17], [18]), suggesting that it is possible to do fine-mapping between breeds. These observations are consistent with two bottlenecks, the first associated with the domestication from wolves and the second associated with the intense selection to create the breeds. Other studies have successfully employed the extensive variation between breeds to map genes affecting size and behavior ([16, 5]). The validity of this approach rests on the assumption that the breeds used are random samples of unrelated breeds or that related breeds make up a small part of our sample ([16, 13]). This is rarely the case and false positive results may therefore have arisen from population structure. A second independent dataset would be useful to confirm our findings. Finally, a functional assay would be necessary to establish if the interactions we found are also biologically meaningful.

ACKNOWLEDGEMENTS We would like to thank Anders Albrechsten, Luke W. Miratrix, Rasmus Nielsen, Lior Pachter, Montgomery Slatkin, Yun S. Song, and Bernd Sturmfels for helpful discussions. We would also like to acknowledge Yu Zhang for his help with BEAM and Heidi Parker and Elaine Ostrander for providing the dog dataset.

References

[1] 4ti2 team. 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at www.4ti2.de.

[2] A. Albrechtsen, S. Castella, G. Andersen, T. Hansen, O. Pedersen, and R. Nielsen. A Bayesian multilocus association method: allowing for higher-order interaction in association studies. *Genetics*, 176(2):1197–1208, 2007.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000.

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

[5] E. Cadieu, M. W. Neff, P. Quignon, K. Walsh, K. Chase, H. G. Parker, B. M. Vonholdt, A. Rhue, A. Boyko, A. Byers, A. Wong, D. S. Mosher, A. G. Elkahloun, T. C. Spady, C. Andre, K. G. Lark, M. Cargill, C. D. Bustamante, R. K. Wayne, and E. A. Ostrander. Coat variation in the domestic dog is governed by variants in three genes. *Science*, 326(5949):150–153, 2009.

[6] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11:2463–2468, 2002.

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

[8] I. Dikic and A. Blaukat. Protein tyrosine kinase-mediated pathways in G protein-coupled receptor signaling. *Cell Biochemistry and Biophysics*, 30(3):369–387, 1999.

[9] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics.* Basel: Birkhüser, Oberwolfach Seminars, Vol. 40, 2009.

[10] M. Emily, T. Mailund, J. Hein, L. Schauser, and M. H. Schierup. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17(10):1231–1240, 2009.

[11] S. E. Fienberg. An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*, 41:907–917, 1970.

[12] W. R. Gilks, S. Richardson, and D. J.(eds.) Spiegelhalter. *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall, 1995.

[13] M. E. Goddard and B. J. Hayes. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet*, 10(6):381–91, 2009.

[14] A. V. Greig, C. Linge, and G. Burnstock. Purinergic receptors are part of a signalling system for proliferation and differentiation in distinct cell lineages in human anagen hair follicles. *Purinergic Signalling*, 4(4):331–338, 2008.

[15] I. B. Hallgrimsdottir and D. S. Yuster. A complete classification of epistatic two-locus models. *BMC Genetics*, 9:17, 2008.

[16] P. Jones, K. Chase, A. Martin, P. Davern, E. A. Ostrander, and K. G. Lark. Single-nucleotide-polymorphism-based association mapping of dog stereotypes. *Genetics*, 179(2):1033–44, 2008.

[17] E. K. Karlsson, I. Baranowska, C. M. Wade, N. H. Salmon Hillbertz, M. C. Zody, N. Anderson, T. M. Biagi, N. Patterson, G. R. Pielberg, E. J. Kulbokas, 3rd, K. E. Comstock, E. T. Keller, J. P. Mesirov, H. von Euler, O. Kampe, A. Hedhammar, E. S. Lander, G. Andersson, L. Andersson, and K. Lindblad-Toh. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet*, 39(11):1321–8, 2007.

[18] K. Lindblad-Toh, C. M. Wade, T. S. Mikkelsen, and E. K. Karlsson. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–19, 2005.

[19] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.

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

[21] K. A. Mather, A. L. Caicedo, N. R. Polato, K. M. Olsen, S. McCouch, and M. D. Purugganan. The extent of linkage disequilibrium in rice (Oryza sativa L.). *Genetics*, 177(4):2223–2232, 2007.

[22] S. A. McCarroll, A. Huett, P. Kuballa, S. D. Chilewski, A. Landry, P. Goyette, M. C. Zody, J. L. Hall, S. R. Brant, J. H. Cho, R. H. Duerr, M. S. Silverberg, K. D. Taylor, J. D. Rioux, D. Altshuler, M. J. Daly, and R. J. Xavier. Deletion polymorphism upstream of irgm associated with altered irgm expression and crohn's disease. *Nature Genetics*, 40:1107–1112, 2008.

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

[24] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, 69:124–137, 2001.

[25] J. M. Smith and J. Haigh. The hitchhiking effect of a favourable gene. *Genetical Research*, 23:23–35, 1974.

[26] N. B. Sutter, M. A. Eberle, H. G. Parker, B. J. Pullar, E. F. Kirkness, L. Kruglyak, and E. A. Ostrander. Extensive and breed-specific linkage disequilibrium in Canis familiaris. *Genome Res*, 14(12):2388–96, 2004.

[27] R. K. Wayne and E. A. Ostrander. Lessons learned from the dog genome. *Trends Genet*, 23(11):557–67, 2007.

[28] N. Weger and T. Schlake. Igf-I signalling controls the hair growth cycle and the differentiation of hair shafts. *Journal of Investigative Dermatology*, 125(5):873–882, 2005.

[29] G. L. Wolff, J. S. Stanley, M. E. Ferguson, P. M. Simpson, M. J. Ronis, and T. M. Badger. Agouti signaling protein stimulates cell division in "viable yellow" (A^{vy}/a) mouse liver. *Experimental Biology and Medicine (Maywood)*, 232(10):1326–1329, 2007.

[30] F. A. Wright, H. Huang, X. Guan, K. Gamiel, C. Jeffries, W. T. Barry, F. Pardo-Manuel de Villena, P. F. Sullivan, K. C. Wilhelmsen, and F. Zou. Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, 23:2581–2588, 2007.

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