LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

A.-L. Boulesteix, C. Strobl, S. Weidinger, H. E. Wichmann & S. Wagenpfeil

Multiple testing for SNP-SNP interactions

Technical Report Number 004, 2007
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Multiple testing for SNP-SNP interactions

A.-L. Boulesteix[*][†], C. Strobl[‡], S. Weidinger[§][¶],
H. E. Wichmann[‖], S. Wagenpfeil[*]

November 8, 2007

Abstract

Most genetic diseases are complex, i.e. associated to combinations of SNPs rather than individual SNPs. In the last few years, this topic has often been addressed in terms of SNP-SNP interaction patterns given as expressions linked by logical operators. Methods for multiple testing in high-dimensional settings can be applied when many SNPs are considered simultaneously. However, another less well-known multiple testing problem arises within a fixed subset of SNPs when the logic expression is chosen optimally. In this article, we propose a general asymptotic approach for deriving the distribution of the maximally selected chi-square statistic in various situations. We show how this result can be used for testing logic expressions -in particular SNP-SNP interaction patterns- while controlling for multiple comparisons. Simulations show that our method provides multiple testing adjustment when the logic expression is chosen such as to maximize the statistic. Its benefit is demonstrated through an application to a real dataset from a large population-based study considering allergy and asthma

[*] Department of Medical Statistics and Epidemiology, Technical University of Munich, Ismaningerstr. 22, 81675 Munich, Germany

[†] Sylvia Lawry Centre, Hohenlindenerstr. 1, D-81677 Munich, Germany

[‡] Department of Statistics, Ludwig-Maximilians-University of Munich, Ludwigstr. 33, 80539 Munich, Germany

[§] Department of Dermatology and Allergology, Technical University of Munich, Biedersteinerstr. 29, 80802 Munich, Germany

[¶] Division of Environmental Dermatology and Allergy TUM/GSF

[‖] Institute for Epidemiology, GSF, Ingolstädter Landstraße, D-85764 Neuherberg, Germany

in KORA. An implementation of our method is available from the Comprehensive R Archive Network (CRAN) as R package 'SNPmaxsel'.

1 Introduction

Meanwhile, there are more than 10 million (year 2005: $>$ 9.2 million according to Consortium (2005)) single nucleotide polymorphisms (SNPs) of the human genome available in public databases. These SNPs are almost all biallelic with two of the four bases A (adenine), C (cytosine), T (thymine) and G (guanine) occurring at the considered locus. From a statistical point of view, a SNP may be seen as a categorical variable with three categories: if the SNP alleles are, e.g., adenine and cytosine, the three possible genotypes are AA, AC and CC. Association studies aim at identifying genetic factors associated with a certain disease. One of the standard approaches to deal with such data is the chi-square test of independence (Sasieni, 1997).

For a given SNP, data can be dichotomized in three different ways: AA vs. $\{AC, CC\}$, CC vs. $\{AA, AC\}$ or AC vs. $\{AA, CC\}$. In general none of these three genetic models for a specific SNP is favored a priori. Performing three tests simultaneously (one for each dichotomization), however, yields a multiple testing problem for which adjustment is needed.

Most diseases are complex, i.e. associated to combinations of SNPs rather than individual SNPs. Hence, many studies focus on associations between a disease and haplotypes rather than or in addition to individual SNP analyses (Becker and Knapp, 2004). A major drawback of such approaches is the uncertainty involved in inferring haplotypes from unphased genotype SNP data and the involvement of a computationally expensive estimation algorithm. Another important approach, which has been given much attention in the last few years and will be considered here, considers combinations of the form, e.g.,

$$(SNP_1 = AA) \;\wedge\; (SNP_2 \in \{TT, TG\}) \qquad (1)$$

that may lead to higher or lower risk of developing a certain disease. In the present article, we denote as a *pattern* a rule like that of Eq. (1) involving one or more conditions of the type $SNP_i \in \mathcal{S}_i$. An interaction pattern is then defined as a pattern involving two or more conditions that are linked with the logical operators $\wedge$ and $\vee$, where i is the index of the considered SNP and $\mathcal{S}_i$ denotes a subset of the three possible genotypes of SNP_i. Searching for such interaction patterns in high-dimensional data is a daunting task. Many search algorithms have been suggested in the context of SNP and gene expression data. Some of them are based on logic regression (Ruczinski et al., 2003) and use a simulated annealing or a Monte-Carlo approach (Ruczinski et al., 2003, 2004; Kooperberg and Ruczinski, 2005; Schwender and Ickstadt, 2007) to search the space of possible interactions. Tree-based methods can also be applied to the search of interactions. They are used by Lunetta et al. (2004); Huang et al. (2004); Bureau et al. (2005) to identify interactions in SNP data, whereas Boulesteix et al. (2003); Boulesteix and Tutz (2006) apply a CART-based algorithm to gene expression data for the same purpose. Nelson et al. (2001) suggest a combinatorial partitioning method. 'Polymorphism interaction analysis' (PIA) is another enumeration-based recently developed method (Goodman et al., 2006). A study involving quantitative traits instead of a disease status can be found in Nelson et al. (2001). For an interesting review, see Hoh and Ott (2003).

Unlike standard logistic regression, all these approaches are not based on additive models. They can typically identify main effects in form of a pattern of order one. However, in the presence of two main effects, they can not tell if there is an additional interaction effect in the sense of linear models. To answer such questions, one may rely on models and predictor selection procedures such as those discussed by Bogdan et al. (2004); Baierl et al. (2006) for the case of quantitative trait loci. Nevertheless, methods based on logical expressions are usually more intuitive to interpret than logistic regression models with interactions, which has probably contributed to their spectacular development observed in the last few years.

Validation and statistical significance of the located (interaction) patterns are delicate issues. Most of the articles mentioned above address the problem in terms of cross-validation error or model size selection. In the present paper, we examine the problem from a completely different point of view: for a given interaction between two variables, we examine the statistical significance in terms of multiple testing, which is not done in the other articles. More precisely, the multiple testing problem occurring in association studies involving interactions can be decomposed into two components:

- If p denotes the number of SNPs in the study, there are $\binom{p}{q}$ ways to choose a subset of q SNPs for constructing a pattern. Correction for multiple testing is essential. It is typically performed via Bonferroni correction (Marchini et al., 2005) or by controlling the False Discovery Rate using the original procedure by Benjamini and Hochberg (1995) or one of its later variants, e.g. Storey (2002). Methods based on the local False Discovery Rate (Efron and Tibshirani, 2002) can also be used in this context. Since the different pairs of SNPs are expected to show strong association, it might be sensible to use a method which is valid for correlated hypotheses (Benjamini and Yekutieli, 2001).

- For a fixed subset of SNPs $\{SNP_1, \dots, SNP_q\}$, there are 3^q possible combinations of genotypes from a combinatorial point of view, yielding $2^{3^q-1} - 1$ possible partitions in binary split analysis. If one considers patterns involving only the operator $\wedge$, the number of possible partitions decreases to 6^q or 4^q if the heterozygous model is not considered. However, even for $q = 2$, this multiple testing component is not negligible. This problem is related to the well-known selection bias occurring in recursive partitioning when the best binary splitting is selected from predictors with different numbers of categories (Kim and Loh, 2001; Boulesteix, 2006a; Strobl et al., 2007). Note that this problem is also crucial in classical methods based on logistic regression, though it is then most often ignored in practice.

In the present article, we focus on the second aspect of the multiple testing issue, which is often ignored in practical studies, and introduce a general framework that can be applied to various problems involving, e.g., interactions between SNPs. One option consists of adjusting the p-values using the maxT procedure based on computationally expensive permutation algorithms. This method is adopted by Sladek et al. (2007) when testing the association of single SNPs with the phenotype in three different models (additive, dominant, recessive). In the present paper, we generalize this idea to interactions and suggest a very fast computation approach which is not based on permutations. More precisely, we derive the asymptotic distribution of the chi-square statistic yielded by the optimal dichotomization of a multicategorical nominal variable. By optimal dichotomization, we mean the dichotomization that yields the highest chi-square statistic out of a set of user-defined candidate dichotomizations. This result is applied to the special case of SNP-SNP interactions.

Our approach can be applied to all types of association studies involving independent patients, including population-based or case-control studies. It is implemented in the R system for statistical computing and freely available as a user-friendly package ('SNPmaxsel') from the Comprehensive R Archive Network (CRAN) at
`http://cran.r-project.org/src/contrib/Descriptions/SNPmaxsel.html`.

2 Approach

This section introduces a new general statistical method related to maximally selected statistics and handling any type of categorical variable. The application to SNP patterns, especially SNP-SNP interactions, is outlined in Section 3.

2.1 Chi-square tests

Let the (unordered) categories of a nominal random variable X (for example a SNP or the pseudo-variable $X_{1,2}$ defined in Section 3) be denoted as $1, \dots, K$. In the case of a SNP, we have $K = 3$, whereas $K = 9$ for $X_{1,2}$ (see Section 3). The two categories of the binary random variable Y (e.g. the disease status) are denoted as $0, 1$. Let p and p_k, for $k = 1, \dots, K$ be defined as $p = P(Y = 1)$ and $p_k = P(Y = 1|X = k)$.

We consider a sample of independent identically distributed observations $(x_i, y_i)_{i=1,\dots,N}$. For $k = 1, \dots, K$ and $c = 0, 1$, we define

$$\begin{aligned} N_{k.} &= \textstyle\sum_{i=1}^{n} I(x_i = k), \\ N_{.c} &= \textstyle\sum_{i=1}^{n} I(y_i = c), \\ n_{kc} &= \textstyle\sum_{i=1}^{n} I(x_i = k) \cdot I(y_i = c). \end{aligned}$$

The sample estimators of p and p_k are then given as $\hat{p} = \frac{N_{.1}}{N}$ and $\hat{p}_k = \frac{n_{k1}}{N_{k.}}$, respectively. More generally, if A is a subset of $\{1, \ldots, K\}$, we define

$$\begin{array}{lcl} p_A & = & P(Y = 1 | X \in A), \\ N_{A.} & = & \sum_{i=1}^{n} I(x_i \in A), \\ n_{Ac} & = & \sum_{i=1}^{n} I(x_i \in A) \cdot I(y_i = c), \\ \hat{p}_A & = & \frac{n_{A1}}{N_{A.}}. \end{array}$$

If B is the complementary set of A, i.e. the set such that $A \cap B = \emptyset$ and $A \cup B = \{1, \ldots, K\}$, the chi-square statistic used to compare p_A and p_B is

$$\chi^2_{A,B} = \frac{N(n_{A1} n_{B0} - n_{A0} n_{B1})^2}{N_{.0} N_{.1} N_{A.} N_{B.}}. \quad (2)$$

It can be easily shown that $\chi^2_{A,B} = Z^2_{A,B}$, where

$$Z_{A,B} = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{N_{A.}} + \frac{1}{N_{B.}} \right)}}. \quad (3)$$

Thus, for any $t \geq 0$, we have

$$\chi^2_{A,B} \leq t \;\Leftrightarrow\; -\sqrt{t} \leq Z_{A,B} \leq \sqrt{t}.$$

It is well-known that $Z_{A,B}$ is asymptotically normally distributed under the null-hypothesis $p_A = p_B$. The chi-square asymptotic test may be performed equivalently based on the chi-square statistic of Eq. (2) or on the Z statistic of Eq. (3).

Suppose we perform m such tests to compare p_{A_i} and p_{B_i}, where $(A_1, B_1), \ldots, (A_m, B_m)$ are pairs of complementary subsets of $\{1, \ldots, K\}$. Note that the theory presented in Section 2.3 can be easily generalized to sets A_i and B_i that are not complementary. Let

$$\chi^2_{max} = \max_{i=1,\ldots,m} \chi^2_{A_i,B_i} \quad (4)$$

define the maximum chi-square statistic. For $t \geq 0$, we have

$$P(\chi^2_{max} \leq t) = P\left(\cap_{i=1}^{m}(-\sqrt{t} \leq Z_{A_i,B_i} \leq \sqrt{t})\right). \tag{5}$$

In Section 2.3, we derive the asymptotic joint distribution of $Z_{A_1,B_1}, \ldots, Z_{A_m,B_m}$ under the null-hypothesis of no association between X and Y, which can be expressed as $H_0: \ p_1 = \cdots = p_K$.

2.2 Connection to the maxT multiple testing procedure

The distribution of χ^2_{max} under the null-hypothesis $H_0: \ p_1 = \cdots = p_K$ may be used to adjust p-values for multiple testing. For $i = 1, \ldots, m$, let $H_0^{(i)}$ denote the null-hypothesis $p_{A_i} = p_{B_i}$ and $p^{(i)}$ the p-value of the corresponding asymptotic chi-square test. H_0 can be written as $H_0 = \cap_{i=1}^{m} H_0^{(i)}$. Correction of the p-values $p^{(1)}, \ldots, p^{(m)}$ for multiple testing may be performed using, e.g., Bonferroni's adjustment procedure. However, in the case of dependent test statistics, the so-called maxT procedure (Westfall and Young, 1993) may be much more powerful (Dudoit et al., 2003). If $t^{(i)}$ denotes the (observed) i-th test statistic, the i-th maxT adjusted p-value $\tilde{p}^{(i)}$ is given as $\tilde{p}^{(i)} = P(\chi^2_{max} \geq t^{(i)}|H_0)$. Interested readers may refer to Dudoit et al. (2003) for more details on adjustment procedures. For large sample sizes, we have $P(\chi^2_{max} \geq t^{(i)}|H_0) = 1 - P(\chi^2_{max} \leq t^{(i)}|H_0)$. The maxT adjusted p-values can be thus computed based on formula (5), using the results described below.

2.3 Using the multivariate normal distribution

In Lausen et al. (2004), the distribution of maximally selected rank statistics over the range of several predictors is approximated as multivariate normal. In the present article, we also use the multivariate normal distribution, but in a different context. We denote the random vector $(Z_{A_1,B_1}, \ldots, Z_{A_m,B_m})^T$ as $\mathbf{z}$, where Z_{A_i,B_i} is defined as in Section 2.1 and derive its multivariate distribution under the null-hypothesis of no association between X and Y, conditional on the marginal counts $N_{k.}$, for $k = 1, \ldots, K$.

Let us consider the random vector $\mathbf{p}$ defined as $\mathbf{p} = (\hat{p}_1, \ldots, \hat{p}_K)^T$. For all pairs of non-empty complementary subsets $A, B \subset \{1, \ldots, K\}$, the numerator of $Z_{A,B}$ can be written as a linear transformation of $\mathbf{p}$:

$$\hat{p}_A = \sum_{k \in A} \frac{N_{k.}}{N_{A.}} \hat{p}_k = \sum_{k \in A} \frac{a_k}{\sum_{j \in A} a_j} \hat{p}_k, \tag{6}$$

where a_k is the proportion of observations in category k defined as $a_k = \frac{N_{k.}}{N}$. Finally, for any sets $(A_1, B_1), \ldots, (A_m, B_m)$, the m-vector $\mathbf{z}$ can be expressed as

$$\mathbf{z} = \frac{\sqrt{N}}{\sqrt{\hat{p}(1-\hat{p})}} \mathbf{A}\mathbf{p}, \tag{7}$$

where $\mathbf{A}$ is the $m \times K$ matrix whose entries depend only on $a_1, \ldots, a_K$. More precisely, the element of $\mathbf{A}$ in the i-th line and j-th column is given as

$$A_{ij} = \begin{cases} \frac{a_j}{\sum_{k \in A_i} a_k} \left(\frac{1}{\sum_{k \in A_i} a_k} + \frac{1}{\sum_{k \in B_i} a_k} \right)^{-1/2} & \text{if } j \in A_i, \\ -\frac{a_j}{\sum_{k \in B_i} a_k} \left(\frac{1}{\sum_{k \in A_i} a_k} + \frac{1}{\sum_{k \in B_i} a_k} \right)^{-1/2} & \text{if } j \in B_i, \\ 0 & \text{else.} \end{cases}$$

For large samples such as those usually considered in association studies, we may use the approximation $\hat{p}(1-\hat{p}) \approx p(1-p)$, like in the classical chi-square test outlined in Section 2.1. Up to this approximation, we have thus

$$\mathbf{z} = \frac{\sqrt{N}}{\sqrt{p(1-p)}} \mathbf{A}\mathbf{p}.$$

Since the term $\frac{\sqrt{N}}{\sqrt{p(1-p)}}$ resulting from the approximation is constant, the covariance matrix $\Sigma_{\mathbf{z}}$ of the random vector $\mathbf{z}$ is given as

$$\Sigma_{\mathbf{z}} = \frac{N}{p(1-p)} \mathbf{A} \Sigma_{\mathbf{p}} \mathbf{A}^T,$$

where $\Sigma_{\mathbf{p}}$ denotes the covariance matrix of $\mathbf{p}$.

The covariance matrix $\Sigma_{\mathbf{p}}$ can be derived as follows. Conditional on the marginal counts $N_{k.}$, the components $\hat{p}_1, \ldots, \hat{p}_K$ of the vector $\mathbf{p}$ are independent. For large $N_{k.}$, $\sqrt{N_{k.}}\hat{p}_k$ $(k = 1, \ldots, K)$ converges to a Gaussian distribution with mean $\sqrt{N_{k.}}p_k$ and variance $p_k(1-p_k)$. The asymptotic covariance matrix $\Sigma_{\mathbf{p}}$ of $\mathbf{p}$ is thus given as

$$\Sigma_{\mathbf{p}} \;=\; diag\left(\tfrac{p_1(1-p_1)}{N_{1.}}, \dots, \tfrac{p_K(1-p_K)}{N_{K.}}\right)$$

and the random vector $\mathbf{p}$ has an asymptotically multivariate normal distribution. Hence, the random vector $\mathbf{z}$ also follows an asymptotically multivariate normal distribution and, under the null-hypothesis $p_1 = \dots = p_K$, its asymptotical covariance matrix is given as

$$\begin{aligned} \Sigma_{\mathbf{z}} &= N\,\frac{p(1-p)}{p(1-p)}\,\mathbf{A}\; diag\left(\frac{1}{N_{1.}}, \dots, \frac{1}{N_{K.}}\right)\mathbf{A}^T & (8) \\ &= \mathbf{A}\; diag\left(\frac{1}{a_1}, \dots, \frac{1}{a_K}\right)\;\mathbf{A}^T. & (9) \end{aligned}$$

In conclusion, under the null-hypothesis, we have the asymptotical result $\mathbf{z} \sim \mathcal{N}_m(\mathbf{0}, \Sigma_{\mathbf{z}})$, with $\Sigma_{\mathbf{z}}$ depending only on $N_{1.}, \dots, N_{K.}$ which are considered as fixed. Based on Eq. (5), the computation of $P_{H_0}(\chi^2_{max} \leq t)$ given $N_{1.}, \dots, N_{K.}$, which involves a multidimensional integral, is then straightforward. It is implemented in our package 'SNPmaxsel', which is available from the Comprehensive R Archive Network (CRAN). This package uses the package 'mvtnorm' (Genz et al., 2006) to compute the multivariate normal distribution function.

3 Evaluating SNP-SNP (interaction) patterns

3.1 Patterns of order one and two

In the present section, we show how the methodology introduced in Section 2 can be applied to the special case of SNP patterns, especially SNP-SNP interactions. We consider the case of two exemplary SNPs X_1 and X_2 with, say, genotypes AA, AC, CC and TT, TG, GG for simplicity, but the theory is essentially generalizable to more SNPs. Let us consider logic expressions involving the $\wedge$ or $\vee$ operators, for instance

$$L = ((X_1 = AA) \wedge (X_2 = TG)) \vee (X_1 = CC).$$

Let us define the random variable $X_{1,2}$ by

$$\begin{aligned}
X_{1,2} &= 1 \text{ if } X_1 = AA \text{ and } X_2 = TT \\
X_{1,2} &= 2 \text{ if } X_1 = AC \text{ and } X_2 = TT \\
X_{1,2} &= 3 \text{ if } X_1 = CC \text{ and } X_2 = TT \\
X_{1,2} &= 4 \text{ if } X_1 = AA \text{ and } X_2 = TG \\
X_{1,2} &= 5 \text{ if } X_1 = AC \text{ and } X_2 = TG \\
X_{1,2} &= 6 \text{ if } X_1 = CC \text{ and } X_2 = TG \\
X_{1,2} &= 7 \text{ if } X_1 = AA \text{ and } X_2 = GG \\
X_{1,2} &= 8 \text{ if } X_1 = AC \text{ and } X_2 = GG \\
X_{1,2} &= 9 \text{ if } X_1 = CC \text{ and } X_2 = GG
\end{aligned} \tag{10}$$

The logical expression L may then be reformulated as $X_{1,2} \in A_L$, with $A_L = \{3, 4, 6, 9\}$.

Many articles in bioinformatics, statistics and genetics are devoted to the search of logic expressions that are linked to the binary variable Y of interest (e.g. disease status). In the context of the chi-square statistic, one would look for a binary partition $\{A_L, \overline{A}_L\}$ maximizing the chi-square statistic obtained for the test of $p_{A_L} = p_{\overline{A}_L}$. There are $2^8 - 1 = 255$ distinct binary partitions of the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Each of them corresponds to several equivalent logic expressions. Conversely, each logic partition leads to a unique partition of $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Considering all the possible partitions of $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ leads to a vector $\mathbf{z}$ of length 255.

However, not all partitions are equally interpretable. From a medical point of view, it makes sense to restrict to logic expressions formed by two terms linked by the $\wedge$ operator, for instance

$$(X_1 = AA) \wedge (X_2 \in \{TT, TG\}).$$

If the heterozygous genotypes are considered as intermediates between the homozygous genotypes, it makes sense to consider only the combination of one of the four terms $X_1 \in \{AA, AC\}$, $X_1 \in \{AC, CC\}$, $X_1 = AA$, $X_1 = CC$ with one of the four terms $X_2 \in \{TT, TG\}$, $X_2 \in \{TG, GG\}$, $X_2 = TT$, $X_2 = GG$ using the $\wedge$ operator, yielding $4 \times 4 = 16$ candidate patterns. The sets (A_i, B_i) $(i = 1, \ldots, 16)$

underlying the vector $\mathbf{z}$ may then be defined as $A_1 = \{1\}, B_1 = \{2, \ldots, 9\}$, $A_2 = \{1, 2\}, B_2 = \{3, \ldots, 9\}$ and so on. Moreover, following the line of logistic regression which incorporates both main and interaction effects, one can also consider patterns of order one, i.e. involving only one term instead of two. This approach yields four additional partitions corresponding to $X_1 = AA$, $X_1 \in \{AA, AC\}$, $X_2 = TT$, $X_2 \in \{TT, TG\}$. For example, $X_1 = AA$ corresponds to the partition $A = \{1, 4, 7\}, B = \{2, 3, 5, 6, 8, 9\}$. Finally, one obtains $m = 16 + 4 = 20$ candidate patterns.

Other variants are conceivable, for instance if one considers that the group of heterozygous genotypes taken alone may be at decreased or increased risk. Although there are some well-known examples of such effects, heterozygous individuals most often have intermediate phenotypes or, as common in monogenic diseases, the same phenotype as the homozygous variant individuals (dominant model) or the homozygous wild-type individuals (recessive model), see Marchini et al. (2005) for an extensive discussion of different plausible genetic models. Hence, restricting to the $4 \times 4 + 4 = 20$ patterns mentioned above often makes sense in practice. However, the extension to other variants is straightforward: one just has to define the desired additional partitions (A_i, B_i) of $\{1, \ldots, 9\}$ or to remove the inadequate partitions. The whole procedure is summarized below.

Algorithm. Scoring a pair of SNPs.

1. Transform the two variables X_1 and X_2 into a single variable $X_{1,2}$ as described in Eq. (10).

2. Compute the maximal chi-square statistic $\chi^{2\,(obs)}_{max}$ obtained by partitioning $X_{1,2}$, using the desired partitions (A_i, B_i), $i = 1, \ldots, m$, for instance the $m = 20$ partitions outlined above.

3. Compute the matrix $\mathbf{A}$ corresponding to the selected partitions of $X_{1,2}$ and the covariance matrix $\Sigma_{\mathbf{z}}$.

4. Derive $P_{H_0}(\chi^2_{max} \leq \chi^{2\,(obs)}_{max})$ from Eq. (5) using the multivariate normal distribution with covariance matrix $\Sigma_{\mathbf{z}}$.

A large value of $P_{H_0}(\chi^2_{max} \leq \chi^{2\ (obs)}_{max})$ indicates a highly discriminating pair of SNPs and $1 - P_{H_0}(\chi^2_{max} \leq \chi^{2\ (obs)}_{max})$ can be seen as an adjusted p-value.

Note that possible associations between SNPs are taken into account through conditioning the distribution on the marginal counts of the variable $X_{1,2}$. In the extreme case of two completely linked SNP loci, the variable $X_{1,2}$ would only take the three values $1, 5, 9$, yielding only two distinct partitions $\{1, 5\}$ vs $\{9\}$, $\{1\}$ vs $\{5, 9\}$. In intermediate cases where the categories other than $1, 5, 9$ have smaller counts, the dependence between the two SNPs is also taken into account since the distribution of χ^2_{max} is derived conditionally on the frequencies of the nine categories.

For two SNPs, each coded as $1, 2, 3$, representing, say, the three genotypes AA, AC, CC, $P_{H_0}(\chi^2_{max} \leq \chi^{2\ (obs)}_{max})$ is computed using the R package 'SNPmaxsel' by

```
>library(SNPmaxsel)
>maxsel.asymp.test(x1=x1,x2=x2,y=y,type="inter.ord.main")
```

for the variant with $m = 20$ described above, where y, x1 and x2 denote the n-vectors giving the value of the response Y, SNP_1 and SNP_2 for the n patients.

Note that, if one is interested in main effects only, better power can be achieved by ignoring interactions and focusing on patterns of order one. We have then $m = 2$ or $m = 3$ comparisons for each SNP, depending whether heterozygotes are allowed to be at lower or higher risk than homozygotes ($m = 3$) or not ($m = 2$). For instance, the case $m = 2$ works as follows. Since we are interested in the main effect of a single SNP, we have $K = 3$. If the three possible genotypes are coded as 1,2,3 (where 2 is the heterozygous genotype), the two candidate patterns correspond to the partitions $A_1 = \{1\}, B_1 = \{2, 3\}$ and $A_2 = \{1, 2\}, B_2 = \{3\}$, respectively. This variant is included in the package 'SNPmaxsel', with `type="ordinal"`, whereas the case $m = 3$ is obtained by setting `type="all.partitions"`. Conversely, if the focus is on interactions only, it might also make sense to consider only the $m = 16$ partitions corresponding to interaction patterns with two involved SNPs. Using 'SNPmaxsel', this can be obtained by setting `type="inter.ord"`.

3.2 Testing multiple patterns simultaneously

A critical issue that may be addressed in future research is the adjustment needed when multiple pairs of SNPs are tested simultaneously. Since these pairs of SNPs may be highly dependent in general, classical adjustment methods assuming independence of the test statistics (e.g., Benjamini and Hochberg, 1995) may be inappropriate. A safe option is to adjust the obtained p-values using a correction procedure that explicitly allows the tested hypotheses to be dependent, for instance the method for controlling the false discovery rate by Benjamini and Yekutieli (2001).

Since such methods are usually too conservative, more research is needed in order to make use of the particular structure of SNP-SNP interaction data. From a theoretical point of view, our method can be extended to the adjustment over several pairs of SNPs. A set of p SNPs can be transformed into a single nominal variable with 3^p categories (of which many will be empty in real data analysis, especially if the considered SNPs are linked). For each of the $p(p-1)/2$ pairs of SNPs, the m hypotheses of interest can be formulated in terms of partitions of this single variable, finally yielding $m' = mp(p-1)/2$ partitions. This approach rises two problems.

Firstly, many cells will have no or few observations. While empty cells are correctly handled by our approach (see Section 4.2), cells with very few observations might alter the results. Hence, efforts should be made to solve this problem, for instance in form of continuity correction. Secondly, the current implementation of the multivariate normal distribution becomes computationally prohibitive for, say, $m > 100$. In further research, one may try to make the algorithm more efficient by using the particular structure of the multiple hypotheses, in the vein of the procedure suggested by Hothorn and Zeileis (2007) for the case of partitions defined by cutpoints.

In the present article, we rely on the conservative method by Benjamini and Yekutieli (2001) for controlling the false discovery rate under dependence of the hypotheses. It is implemented, e.g., in the R package 'multtest'.

3.3 Higher-order interactions

From a theoretical point of view, our method is also generalizable to more complex interaction patterns involving three or more variables. For interactions involving three SNPs, SNP_1, SNP_2 and SNP_3 have to be transformed into a variable $X_{1,2,3}$ with 27 categories using the same scheme as in Eq. (10). Let us consider, as in Section 3.1, the important case of patterns which i) consist of conditions linked by the $\wedge$ operator, ii) do not involve the heterozygous genotype taken alone. Under these restrictions, there are $4 \times 4 \times 4 = 64$ possible patterns of order three. This is because, for each of the three SNPs, there are four possible ways to define the condition. Similarly, there are 16 patterns of order two for each of the three SNP pairs (SNP_1/SNP_2, SNP_1/SNP_3, SNP_2/SNP_3) but only two possible main effects for each of the three SNPs. Finally, it results into $m = 64 + 16 \times 3 + 2 \times 3 = 118$ possible patterns, i.e. 118 partitions (A_i, B_i).

Using the current implementation of the multivariate normal distribution, the distribution of χ^2_{max} in this setting can hardly be determined in reasonable time. However, as discussed in Section 4.2, more efficient algorithms making use of the data structure could be developed in the near future in the vein of the procedure suggested by Hothorn and Zeileis (2007). Alternatives are (likewise computationally intensive) permutation procedures or simulations based on samples drawn from the multivariate normal distribution. However, limited power is to expect, considering the high number of tested hypotheses and the small size of many cells of the three-dimensional contingency table. Note that this lack of power should rather be seen a consequence of the data structure than as an inconvenience of our method: all methods addressing the statistical significance of higher-order patterns face this problem.

4 Simulation study

4.1 Power study

The aim of this simulation is to demonstrate that our method performs adjustment as expected (with control of the type I error) and to compare its power to the power of competing approaches in various settings. For each setting, some 10000 data sets of sample size N=1000 including a binary variable Y and two SNPs SNP_1 and SNP_2 are simulated as follows. For both SNP_1 and SNP_2, the Hardy-Weinberg equilibrium is assumed and different allele frequencies are implemented by means of the parameters π_1 and π_2 indicating the frequency of the more frequent allele for SNP_1 and SNP_2, respectively. Therefore, for $j = 1, 2$, the three genotypes for SNP_j (AA, Aa and aa) are sampled with probabilities π_j^2, $2\pi_j(1 - \pi_j)$ and $(1 - \pi_j)^2$, respectively. SNP_1 and SNP_2 are mutually independent. The parameters (π_1, π_2) are set successively to $(0.6, 0.6)$, $(0.6, 0.8)$, $(0.8, 0.8)$ and $(0.6, 0.95)$. For the distribution of the binary variable Y, we examine three different cases.

a) The binary response Y is sampled independently of SNP_1 and SNP_2 with the marginal class probability $p = 0.5$ (corresponding to balanced case-control studies).

b) The probability $p = P(Y = 1)$ is 0.7 for patients with $(SNP_1 = aa) \wedge (SNP_2 = aa)$, 0.3 for the other. This case corresponds to an interaction pattern of the form of those simulated in Schwender and Ickstadt (2007).

c) The probability $P(Y = 1)$ is 0.55 for patients with $(SNP_1 = aa)$, 0.45 for the other. This case corresponds to a main effect with threshold (recessive model), which can be seen as a pattern of order one. Note that, when there are only main effects, better power could be achieved by focusing on patterns of order one (see end of Section 3). Through the case c), we solely aim to show how the method behaves in the case where there is only a main effect.

Seven assessment approaches are compared:

1) **naive**: The maximally selected chi-square statistic is calculated. It is then referred to the nominal chi-square distribution with one degree of freedom to yield the p-value of the corresponding asymptotic chi-square test.

2) **naive Bonf**: The p-value derived in 1) is adjusted using Bonferroni's method, i.e., by multiplying it through the number of chi-square tests (here, $m = 20$). Other adjustment methods such as Sidak's adjustment could be applied, but usually yield similar results (see Dudoit et al. (2003) for an overview).

3) **chisq8**: the p-value of the chi-square test for independence with eight degrees of freedom obtained after transforming the two SNPs into a single categorical variable with nine classes.

4) **new**: The adjusted p-value is derived using our novel method with $m = 20$ corresponding to 16 interaction patterns and 4 main effects.

5) **log raw**: The standard approach to assess the association between a pair of SNPs and a response variable in genetic association studies consists of building various logistic regression models involving one or both SNP(s), with and without interaction terms (Marchini et al., 2005; Park and Hastie, 2007). Typically, dummy coding is carried out: each SNP variable SNP_j $(j = 1, 2)$ is recoded as two binary variables SNP_j^A and SNP_j^B (corresponding, e.g., to the recessive and dominant models, respectively), yielding a total of $2 \times 2 = 4$ variables. For fair comparison with our approach which assesses both interaction patterns and main effects, we build the following nested logistic models: the four models obtained with either SNP_1^A, SNP_1^B, SNP_2^A or SNP_2^B as single covariate, the four models including a coding from SNP_1 and a coding from SNP_2 without interaction, and the four models including a coding from SNP_1 and a coding from SNP_2 with interaction term. The model yielding the smallest p-value with the likelihood ratio test (against the null model, for fair comparison) is selected, because the underlying genetic model (recessive, dominant, etc) is in general unknown for new complex diseases.

6) **log Bonf**: The p-value selected in 4) is Bonferroni adjusted (by multiplying it through the number of tests $m = 12$).

7) **log global**: The p-value of the likelihood ratio (LR) test with eight degrees of freedom in the model with two SNPs as qualitative predictors and their interactions.

The percentage of p-values that go below the 5%-level is displayed in Table 1 for the seven methods and for all settings. In the null case, both the naive approach and the logistic regression with minimally selected p-values misleadingly

		allele probabilities (π_1, π_2)			
Type		(0.6,0.6)	(0.6,0.8)	(0.8,0.8)	(0.6,0.95)
a)	naive	0.427	0.418	0.395	0.312
null	naive Bonf	0.033	0.029	0.029	0.018
	chisq8	0.049	0.047	0.046	0.038
	new	**0.046**	**0.045**	**0.047**	**0.039**
	log raw	0.235	0.247	0.242	0.258
	log Bonf	0.026	0.027	0.028	0.021
	log global	0.054	0.056	0.061	0.055
b)	naive Bonf	0.893	0.259	0.049	0.023
inter-	chisq8	0.875	0.300	0.085	0.049
action	**new**	**0.911**	**0.309**	**0.078**	**0.047**
	log Bonf	0.774	0.159	0.038	0.024
	log global	0.843	0.278	0.105	0.061
c)	naive Bonf	0.345	0.333	0.087	0.302
main	chisq8	0.326	0.323	0.099	0.337
	new	**0.400**	**0.393**	**0.121**	**0.398**
	log Bonf	0.343	0.341	0.089	0.326
	log global	0.327	0.339	0.124	0.392

Table 1: Percentage of p-values that go below the 5%-level for the seven approaches and the three cases: null case (a, top), interaction (b, middle) and main effect (c, bottom). Different allele probabilities (π_1, π_2) are considered.

produce p-values that go below the 5%-level for the type I error in up to 42 % of the cases – a fact that would lead to severe misinterpretations of the results in practice. Bonferroni adjusted p-values yield type I error rates below the 5%-level. In contrast, our new approach, the chi-square test with eight degrees of freedom and the global LR test roughly hold the 5%-level for the type I error in almost all cases. In the case of the LR test, the type I error rate slightly exceeds the 5% level, but the difference is not statistically significant.

For cases b) and c), we consider only those methods which correctly control the type I error. It turns out that our new adjustment approach improves the power noticeably compared to the Bonferroni adjustment. In terms of power, our approach is similar to both the chi-square test with eight degrees of freedom and the global LR test in some cases (with $(0.8, 0.8)$ and $(0.6, 0.95)$), but outperforms them noticeably in other cases (with $(0.6, 0.6)$, $(0.6, 0.8)$).

Unsurprisingly, the power increases considerably with the number of observations in the cell $SNP_1 = aa \wedge SNP_2 = aa$ for all approaches. For smaller sample sizes we again find comparable results for the allele probabilities π_1 and π_2 that do not produce extremely sparse cell counts.

Another interesting feature of the approach based on multiple chi-square tests is its ability to identify the right pattern (independently of the obtained p-value). Note that this aspect is not related to our adjustment procedure and is ignored by methods such as the chi-square test with eight degrees of freedom or logistic regression approaches. To address this question, we compute the proportion of iterations in which the multiple chi-square tests identified $(SNP_1 = aa) \wedge (SNP_2 = aa)$ as best pattern (i.e. as the pattern with the highest chi-square statistic). Unsurprisingly, this proportion is low for the allele frequencies $(0.8, 0.8)$ and $(0.6, 0.95)$ yielding the smallest risk class (23% and 49%, respectively). The success rate is higher (55%) for $(0.6, 0.8)$ and almost maximal (94%) for $(0.6, 0.6)$. In case of failure, the identified pattern is most often very similar to the right pattern, e.g., $(SNP_1 = aa) \wedge (SNP_2 \in \{Aa, aa\})$.

4.2 Small samples

For each of the four combinations of allele probabilities (π_1, π_2) (see section 4.1) and different sample sizes N ($N = 100, 200, 300, 500, 800, 1000, 1500, 3000, 5000$), we generate 20 data sets consisting of two SNPs. That is, we draw randomly two (mutually independent) SNP variables with three categories, where the probabilities of the three categories are given by $(\pi_1^2, 2\pi_1(1-\pi_1), (1-\pi_1)^2)$ and $(\pi_2^2, 2\pi_2(1-\pi_2), (1-\pi_2)^2)$, respectively. Each data set yields its own configuration of marginal frequencies $N_{1.}, \dots, N_{9.}$ for the categorical variable $X_{1,2}$ derived from the two SNPs (see section 3.1).

For each of these 20 configurations of marginal frequencies $N_{1.}, \dots, N_{9.}$, we simulate $B = 500$ independent binary variables Y with $P(Y = 1) = 0.5$. We compare the theoretical conditional distribution of χ^2_{max} yielded by our method to the empirical distribution obtained via the $B = 500$ iterations based on the Kolmogorov-Smirnov (K-S) statistic for one sample. The average K-S statistic over the 20 configurations is displayed in Figure 1 against N for the four combinations allele probabilities (π_1, π_2) considered in Section 4.1. Unsurprisingly, our approximation is better for allele frequencies yielding large cell counts. An interesting feature is the peak observed for $(0.6, 0.95)$ at $N = 800$, which can be explained as follows. Our method can deal with empty cells of the contingency tables by eliminating the corresponding partitions from the sets of partitions (A_i, B_i), $i = 1, \dots, m$. Hence, the quality of the approximation is not affected by empty cells. In contrast, non-empty cells with few patients are more problematic. For the allele frequencies $(0.6, 0.95)$, empty cells are very likely for small sample sizes, which explains the relatively good quality of the approximation. For increasing N, cells with few observations replace empty cells, thus making the approximation worse. The approximation then improves for very large values of N.

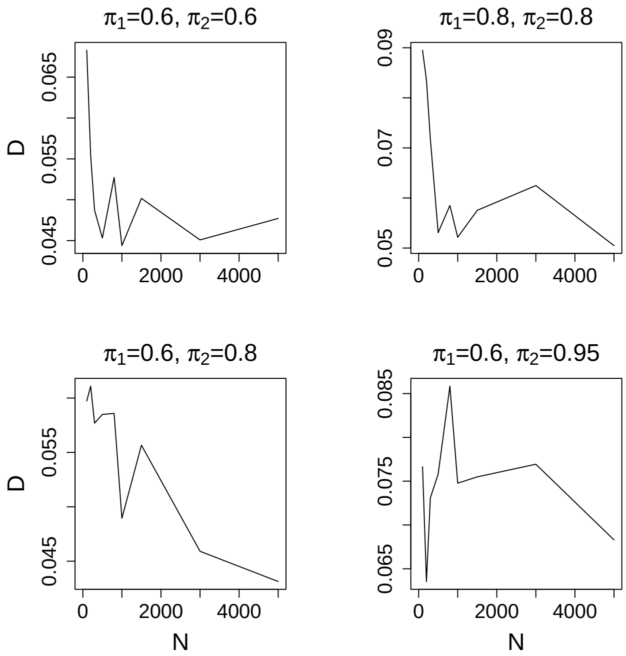

Figure 1: Average K-S statistic against N for $\pi_1 = \pi_2 = 0.6$ (top left), $\pi_1 = \pi_2 = 0.8$ (top right), $\pi_1 = 0.6, \pi_2 = 0.8$ (bottom left), $\pi_1 = 0.6, \pi_2 = 0.95$ (bottom right). Averaged over 20 data sets.

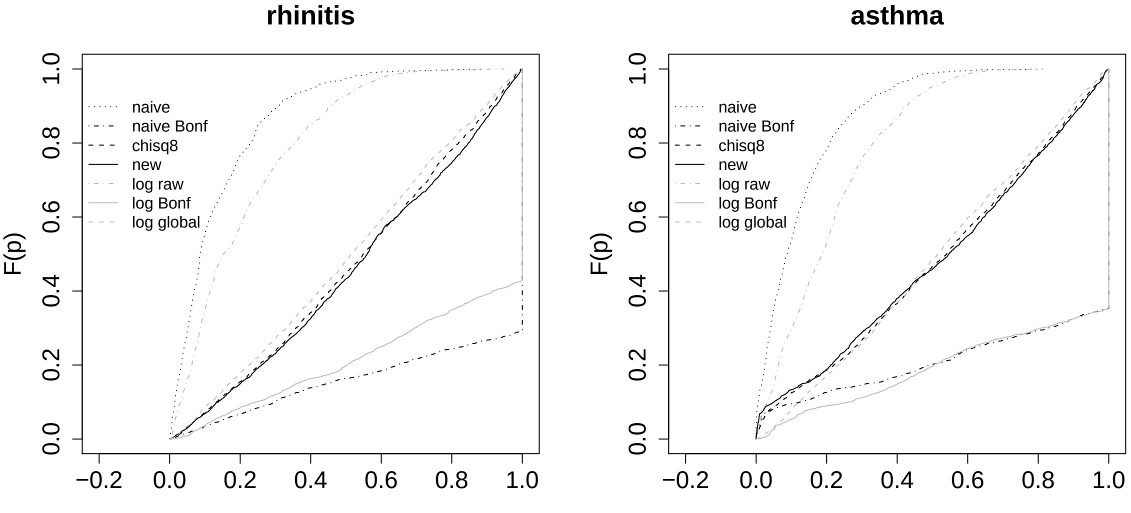

Figure 2: Empirical distribution function $F(p)$ of the p-values yielded by the seven methods described in Section 4.1 for the $66 \times 65/2 = 2145$ SNP-SNP interactions with allergic rhinitis outcome (left) and asthma outcome (right).

5 KORA data example

5.1 The KORA study

KORA (Cooperative Health Research in the Region of Augsburg) is the framework for large population-based surveys on adults in the city and region of Augsburg, South Germany (Holle et al., 2005; Wichmann et al., 2005). The KORA C survey is based on a random, cross-sectional sample stratified for age and sex studied in the years 1994 and 1995. The sampling strategy and study design have been described elsewhere (Weidinger et al. (2004, 2005a,b,c); Illig et al. (2003)). For 1537 individuals, 66 SNPs from five different genes (among others CARD15, STAT6, NOD1) are available. The binary outcomes of interest are 'allergic rhinitis' (388 diseased patients, the rest healthy) and 'asthma' (136 diseased patients, the rest healthy). Considering only pairwise interactions, there are $66 \times 65/2 = 2145$ pairs of SNP-variables.

5.2 Results

We analyze these 2145 pairs of SNPs using all seven methods outlined in Section 4.1. For both examined binary outcomes ('allergic rhinitis' and 'asthma') and both approaches (chi-square tests and logistic regression), it can be seen from the concave curves displayed in Figure 2 that small raw p-values are noticeably more represented than large raw p-values, indicating possible interactions. On the contrary, large p-values are more frequent than small p-values for both approaches when Bonferroni adjustment is carried out. In particular, a large proportion of adjusted p-values equals one. In contrast, our method, the chi-square test with eight degrees of freedom and the global LR test produce p-values that are approximately uniformly distributed within $[0, 1]$, with a slight inflation near zero.

For the outcome 'allergic rhinitis', our new approach yields 62 p-values smaller than 0.05, which are all greater than 0.001. For the outcome 'asthma', there are 30 significant p-values at the level 0.001, and 175 at the level 0.05. However, these p-values have to be adjusted for multiple testing, since 2145 tests are performed simultaneously.

5.3 Adjustment for multiple testing

Adjustment for multiple testing is performed using the stepwise procedure by Benjamini and Yekutieli (2001), which controls the false discovery rate (Benjamini and Hochberg, 1995) under dependence of the tested hypotheses.

After adjustment of the p-values obtained with our new method, all p-values are insignificant at the level 0.05 for the outcome 'allergic rhinitis'. In contrast, the analysis of the outcome 'asthma' yields ten slightly significant p-values (all larger than 0.01). From these ten significant p-values, five involve the SNP dhrs5250, which shows a main effect corresponding to the pattern

$$\text{dhrs5250} = aa,$$

yielding a raw p-value of 0.0029 with the chi-square test. Pairwise interactions patterns are observed between dhrs5250 and the SNPs dhrs4396, dhrs2067, dhrs2064, dhrs1159, dhrs2062, which all show no significant main effect. The five remaining patterns are formed only by SNPs showing no main effect. It should be noted that the ten identified interaction patterns are scarcely significant and based on relatively few diseased patients (136 for the outcome asthma) -as common in population-based studies. Hence, validation will be crucial.

An important point regarding the outcome 'asthma' is that one would obtain 29 instead of 10 significant interaction patterns if the p-values were not adjusted using our new method, thus yielding far too optimistic results in terms of power. Hence, this real data study confirms drastically the results of the simulation study, namely that omitting the p-value adjustment within a given pair of SNPs may lead to over-optimistic conclusions.

6 Conclusion

We have proposed a method to score SNP-SNP (interaction) patterns in association studies while correcting for optimal selection effects. Our approach is based on the derivation of the asymptotic distribution of the maximally selected chi-square statistic in a general context based on the multivariate normal distribution. The simulation results have shown drastically that such an adjustment is necessary, since raw p-values lead to far too high type I errors. In terms of power, our method surpasses logistic regression combined with Bonferroni correction and also often outperforms the chi-square test with eight degrees of freedom and the global LR test for the logistic model, when the true data generating model involves interaction patterns. Our procedure is flexible and can be derived in several variants, depending on the tested hypotheses. For example, the variant used here in the applications considers patterns of order one (corresponding to main effects with threshold) and two.

Furthermore, the obtained distribution is conditional on the marginal frequencies of the genotypes, thus taking associations between SNPs into account. From a theoretical point of view, our approach is related to the conditional inference framework for permutation tests reviewed in Hothorn et al. (2006). In future research, one could attempt to adapt our asymptotic method for scoring interactions to this general framework.

Since based on asymptotic approximations, our procedure can be applied to large samples only. Even if the sample is large, sparse cell counts may occur, for instance for interactions involving the homozygous mutant. As a rule of thumb, the method may be applied if the data approximately fulfill the conditions required by the usual chi-square test. In future research, continuous correction procedures or exact methods based on the methodologies presented in Boulesteix (2006a,b); Boulesteix and Strobl (2007) could potentially be developed in order to solve (at least partly) the problem of sparse cell counts. Permutation-based procedures are another option. However, even exact or permutation-based methods can not tell us whether a pattern present in, say, only two of 1000 patients is significantly associated with a certain disease. From a statistical point of view, improving the sample size is then the only option.

Our method addresses the problem of multiple tests implied by choosing the logical expression optimally for a given pair of SNPs. To the best of our knowledge, this issue is ignored by recent methods based on logical expressions and inspired from machine learning. We think that the concept of logical expressions has advantages over traditional approaches based on (generalized) linear models and shows considerable promise. However, the assessment of statistical significance in this context should be given more attention than usually done. Our method, which is flexible enough to adapt to different statistical questions, may be seen as a contribution to this arduous topic.

Acknowledgement

This work was supported by a grant from the German Ministry of Education and Research (BMBF)/National Genome Research Network (NGFN), project NUW-S31T05. SW and SW were supported by research grants KKF-07/04 and KKF-27/05 of the University Hospital 'Rechts der Isar', Technical University Munich. CS was supported by the German Research Foundation (SFB 386, 'Statistical Analysis of Discrete Structures'). ALB was supported by the Porticus Foundation in the context of the International School for Technical Medicine and Clinical Bioinformatics. We thank Christine Porzelius for helpful comments.

References

Baierl, A., Bogdan, M., Frommlet, F., Futschik, A., 2006. On locating multiple interacting quantitative trait loci in intercross designs. Genetics 173, 1693–703.

Becker, T., Knapp, M., 2004. A powerful strategy to account for multiple testing in the context of haplotype analysis. American Journal of Human Genetics 75, 561–570.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29, 1165–1188.

Bogdan, M., Ghosh, J. K., Doerge, R. W., 2004. Modifying the schwarz bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics 167, 989–999.

Boulesteix, A.-L., 2006a. Maximally selected chi-square statistics and binary splits of nominal variables. Biometrical Journal 48, 838–848.

Boulesteix, A. L., 2006b. Maximally selected chi-square statistics for ordinal variables. Biometrical Journal 48, 451–462.

Boulesteix, A.-L., Strobl, C., 2007. Maximally selected chi-square statistics and non-monotonic associations: an exact approach based on two cutpoints. Computational Statistics and Data Analysis 51, 6295–6306.

Boulesteix, A. L., Tutz, G., 2006. Identification of interaction patterns and classification with applications to microarray data. Computational Statistics and Data Analysis 50, 783–802.

Boulesteix, A. L., Tutz, G., Strimmer, K., 2003. A CART-based method to discover emerging patterns in microarray data. Bioinformatics 19, 2465–2472.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., Eerdewegh, P. V., 2005. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology 28, 171–182.

Consortium, T. I. H., 2005. A haplotype map of the human genome. Nature 437, 1299–1320.

Dudoit, S., Shaffer, J. P., Boldrick, J. C., 2003. Multiple hypothesis testing in microarray experiments. Statistical Science 18, 71–103.

Efron, B., Tibshirani, R., 2002. Empirical bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23, 70–86.

Genz, A., Bretz, F., Hothorn, T., 2006. The mvtnorm package. R package version 0.7-5.
URL `http://CRAN.R-project.org/`

Goodman, J., Mechanic, L., Luke, B., Ambs, S., Chanock, S., Harris, C., 2006. Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis. International Journal of Cancer 118, 1790–1797.

Hoh, J., Ott, J., 2003. Mathematical multi-locus approaches localizing complex human trait genes. Nature Reviews 4, 701–709.

Holle, R., Happich, M., Lowel, H., Wichmann, H., Group, M. S., 2005. Kora–a research platform for population based health research. Gesundheitswesen 67, 19–25.

Hothorn, T., Hornik, K., van de Wiel, M. A., Zeileis, A., 2006. A lego system for conditional inference. The American Statistician 60, 257–263.

Hothorn, T., Zeileis, A., 2007. Generalized maximally selected statistics. Technical Report. Research Report Series / Department of Statistics and Mathematics, Nr. 52. Wien, Wirtschaftsuniv.

Huang, J., Lin, A., Narasimhan, B., T.Quertermous, Hsiung, C. A., Ho, L. T., Grove, J. S., Olivier, M., Ranade, K., Risch, N. J., Olshen, R. A., 2004. Tree-structured supervised learning and the genetics of hypertension. Proceedings of the National Academy of Science 101, 10529–10534.

Illig, T., Bongardt, F., Schöpfer, A., Holler, R., Müller, S., Rathmann, W., et al., Group, K. S., 2003. The endotoxin receptor tlr4 polymorphism is not associated with diabetes or components of the metabolic syndrome. Diabetes 52, 2861–2864.

Kim, H., Loh, W., 2001. Classification trees with unbiased multiway splits. Journal of the American Statistical Association 96, 589–604.

Kooperberg, C., Ruczinski, I., 2005. Identifying interacting SNPs using Monte Carlo logic regression. Genetic Epidemiology 28, 157–170.

Lausen, B., Hothorn, T., Bretz, F., Schumacher, M., 2004. Assessment of optimal selected prognostic factors. Biometrical Journal 46, 364–374.

Lunetta, K. L., Hayward, L. B., Segal, J., Eerdewegh, P. V., 2004. Screening large-scale association study data: exploiting interactions using random forests. BMC Genetics 5, 32.

Marchini, J., Donelly, P., Cardon, L. R., 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics 37, 413–417.

Nelson, M., Kardia, S., Ferrell, R., Sing, C., 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Research 11, 458–470.

Park, M. Y., Hastie, T., 2007. Penalized logistic regression for detecting gene interactions. Biostatistics, (accepted).

Ruczinski, I., Kooperberg, C., LeBlanc, M., 2003. Logic regression. Journal of Computational and Graphical Statistics 12, 475–511.

Ruczinski, I., Kooperberg, C., LeBlanc, M., 2004. Exploring interactions in high dimensional genomic data: an overview of logic regression, with applications. Journal of Multivariate Analysis 90, 178–195.

Sasieni, H. D., 1997. From genotypes to genes: doubling the sample size. Biometrics 53, 1253–1261.

Schwender, H., Ickstadt, K., 2007. Identification of SNP interactions using logic regression. Biostatistics doi:10.1093/biostatistics/kxm024.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C., Froguel, P., 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445, 828–830.

Storey, J., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society B 64, 479–498.

Strobl, C., Boulesteix, A. L., Augustin, T., 2007. Unbiased split selection for classification trees based on the Gini Index. Computational Statistics and Data Analysis 52, 483–501.

Weidinger, S., Klopp, N., Rümmler, L., Wagenpfeil, S., Baurecht, H., Gauger, A., Darsow, U., Jakob, T., Novak, N., Schäfer, T., Heinrich, J., Behrendt, H., Wichmann, H., Ring, J., Illig, T., 2005a. Assocation of CARD15 polymorphisms with atopy-related traits in a population-based cohort of Caucasian adults. Clinical and Experimental Allergy 35, 866–872.

Weidinger, S., Klopp, N., Rümmler, L., Wagenpfeil, S., Novak, N., Baurecht, H., Groer, W., Darsow, U., Heinrich, J., Gauger, A., Schäfer, T., Jakob, T., Behrendt, H., Wichmann, H., Ring, J., Illig, T., 2005b. Assocation of NOD1 polymorphisms with atopic eczema and related phenotypes. Journal of Allergy and Clinical Immunology 116, 177–184.

Weidinger, S., Klopp, N., Wagenpfeil, S., Rümmler, L., Schedel, M., Kabesch, M., Schäfer, T., Darsow, U., Jakob, T., Behrendt, H., Wichmann, H., Ring, J., Illig, T., 2004. Assocation of a STAT6 haplotype with elevated serum IgE levels in a population based cohort of white adults. Journal of Medical Genetics 41, 658–663.

Weidinger, S., Rümmler, L., Klopp, N., Wagenpfeil, S., Baurecht, H., Fischer, G., Holle, R., Gauger, A., Schäfer, T., Jakob, T., Ollert, M., Behrendt, H., Wichmann, H., Ring, J., Illig, T., 2005c. Assocation study of mast cell chymase polymorphisms with atopy. Allergy 60, 1256–1261.

Westfall, P. H., Young, S. S., 1993. Resampling-Based Multiple Testing. Wiley, New York.

Wichmann, H.-E., Gieger, C., Illig, T., 2005. Kora-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67, 26–30.