How to analyze many contingency tables simultaneously in genetic association studies

Dickhaus, Thorsten ; Straßburger, Klaus ; Schunk, Daniel ; Morcillo-Suarez, Carlos ; Illig, Thomas ; Navarro, Arcadi

Abstract: We study exact tests for (2 x 2) and (2 x 3) contingency tables, in particular exact chi-squared tests and exact tests of Fisher type. In practice, these tests are typically carried out without randomization, leading to reproducible results but not exhausting the significance level. We discuss that this can lead to methodological and practical issues in a multiple testing framework when many tables are simultaneously under consideration as in genetic association studies.Realized randomized p-values are proposed as a solution which is especially useful for data-adaptive (plug-in) procedures. These p-values allow to estimate the proportion of true null hypotheses much more accurately than their non-randomized counterparts. Moreover, we address the problem of positively correlated p-values for association by considering techniques to reduce multiplicity by estimating the "effective number of tests" from the correlation structure.An algorithm is provided that bundles all these aspects, efficient computer implementations are made available, a small-scale simulation study is presented and two real data examples are shown

DOI: https://doi.org/10.1515/1544-6115.1776

Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 4 2012 *Article* 12

How to analyze many contingency tables simultaneously in genetic association studies

Thorsten Dickhaus, *Humboldt-University, Berlin*
Klaus Straßburger, *German Diabetes Center, Düsseldorf*
Daniel Schunk, *Johannes Gutenberg-Universität Mainz and University of Zurich*
Carlos Morcillo-Suarez, *Universitat Pompeu Fabra, Barcelona*
Thomas Illig, *Helmholtz Zentrum München*
Arcadi Navarro, *ICREA and Universitat Pompeu Fabra, Barcelona*

Recommended Citation:
Dickhaus, Thorsten; Straßburger, Klaus; Schunk, Daniel; Morcillo-Suarez, Carlos; Illig, Thomas; and Navarro, Arcadi (2012) "How to analyze many contingency tables simultaneously in genetic association studies," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 4, Article 12.
DOI: 10.1515/1544-6115.1776

How to analyze many contingency tables simultaneously in genetic association studies

Thorsten Dickhaus, Klaus Straßburger, Daniel Schunk, Carlos Morcillo-Suarez, Thomas Illig, and Arcadi Navarro

Abstract

We study exact tests for (2 x 2) and (2 x 3) contingency tables, in particular exact chi-squared tests and exact tests of Fisher type. In practice, these tests are typically carried out without randomization, leading to reproducible results but not exhausting the significance level. We discuss that this can lead to methodological and practical issues in a multiple testing framework when many tables are simultaneously under consideration as in genetic association studies.

Realized randomized p-values are proposed as a solution which is especially useful for data-adaptive (plug-in) procedures. These p-values allow to estimate the proportion of true null hypotheses much more accurately than their non-randomized counterparts. Moreover, we address the problem of positively correlated p-values for association by considering techniques to reduce multiplicity by estimating the "effective number of tests" from the correlation structure.

An algorithm is provided that bundles all these aspects, efficient computer implementations are made available, a small-scale simulation study is presented and two real data examples are shown.

KEYWORDS: contingency tables, effective number of tests, genome-wide association study, multiplicity correction, realized randomized p-values, validation stage

Author Notes: This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the Wellcome Trust Case Control Consortium project was provided by the Wellcome Trust under award 076113. The authors like to thank two anonymous referees for their constructive comments which helped to improve the manuscript. Special thanks are due to Prof. Shili Lin for her expeditious handling of all manuscript versions.

1 Introduction

Statistical inference in contingency tables is ubiquitous in genetic association analyses. In particular, depending on the hypothesized underlying genetic model, an analysis of the association between a dichotomous endpoint (like the diagnosis of a disease) and a bi-allelic set of potentially predictive genetic markers can be formalized statistically by a family of tests for association in (2 x 2) or (2 x 3) contingency tables. For a more detailed discussion of the appropriate choice of table layout according to genetic modeling, see, for instance, Chapter 10 in the textbook by Ziegler and König (2006).

Although the theory of exact tests for contingency table analyses can be traced back to Fisher (1922), it continues to pose a challenge for researchers today. Among other things, this is due to unexpected but very interesting phenomena originating from the discreteness of the testing problem. For instance, Finner and Straßburger (2001a,b) showed that the power of contingency table-based tests for association is not monotonic in the sample size. Furthermore, discrete tests are typically carried out without randomization in practice, ensuring reproducible test results but not exhausting the significance level. While this is acceptable for a single comparison, it becomes a serious issue if many contingency tables rather than a single one must be considered simultaneously, as has frequently been done in recent studies. In the latter case, multiplicity correction arises as a further difficulty.

In this article, we will first demonstrate that the performance (in terms of multiple power that we define formally at the end of Section 2) of many modern data-adaptive plug-in multiple tests deteriorates dramatically when discretely distributed p-values are used. Then, we will propose a convenient remedy, focusing on a specific setting for an association study throughout our work: we assume that all markers with alleles that have been successfully identified (i. e., genotyped) will be evaluated simultaneously with respect to their association with a dichotomous phenotype in a confirmatory analysis (no further independent replication study, strong control of the family-wise error rate). Moreover, we assume that the study may consist of two stages: A screening stage and a validation stage, with independent data. From the statistical perspective, this two-stage approach has already been described in detail by Wasserman and Roeder (2009) and Meinshausen et al. (2009). The present article proposes several improvements in statistical inference methods for contingency table analyses under this setting.

In genetic association studies, binary single nucleotide polymorphisms (SNPs) are typically used as genetic markers. Our proposed methodology can be applied to SNP studies, but is also suitable for treating more complex markers such as copy number variations (CNVs) of sections of the deoxyribonucleic acid, as long as the CNVs have the same binary status as SNPs as considered by McCarroll et al.

(2008), for example.

The paper is organized as follows. We will briefly describe classical methods for contingency table analyses under the assumptions mentioned above in Section 2. Experienced readers may skip this section, because it has mainly repetitive character. Section 3 will then present our main contributions: first, we propose various ways to improve the classical strategies while still maintaining tight FWER control; then, we discuss a new algorithm that bundles these approaches. The behavior of the new algorithm in the case of small systems of hypotheses will be investigated by means of Monte Carlo-simulations in Section 4. Details on the necessary computational steps, on numerical feasibility and on resource-efficient implementations will be given in Section 5. Section 6 is devoted to applications of the new method to real-life data sets for type II diabetes and Crohn's disease. We conclude with a discussion in Section 7.

2 Classical approaches

2.1 Notational setup

In what follows, M denotes the number of considered markers. Note that markers can be both genotyped (observed) or imputed (i. e., estimated using population genetics techniques and a priori information from a reference population, see Marchini et al. (2007), Willer et al., 2008). Imputed marker genotypes usually have a very high degree of certainty, so they are widely considered as regular observed genotypes (cf. Howie et al. (2009), Li et al. (2010), The 1000 Genomes Consortium, 2010). We assume that the two rows of the tables under consideration correspond to the phenotype (typically, the disease status) and their (two or three) columns contain the marker counts. Since we want to treat the cases of (2×2) and (2×3) tables simultaneously all along the way, we will denote by $\mathbf{n}$ the vector containing all the (given) marginals of the table. Therefore, $\mathbf{n}$ can have different dimensionality depending on the context. In the (2×2) table case, we have $\mathbf{n} = (n_{1.}, n_{2.}, n_{.1}, n_{.2}) \in \mathbb{N}^4$ while we have $\mathbf{n} = (n_{1.}, n_{2.}, n_{.1}, n_{.2}, n_{.3}) \in \mathbb{N}^5$ in the (2×3) table case. In both cases, we define the number of observational units by $N = n_{1.} + n_{2.}$. In the case of a (2×3) table, N is therefore equal to the number of individuals in the study, while it equals the number of alleles (twice the number of study participants) in the case of a (2×2) table. Accordingly, an observed table will be denoted by $\mathbf{x}$ taking the form $\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \in \mathbb{N}^{2\times 2}$ in case of a (2×2) table and $\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{pmatrix} \in \mathbb{N}^{2\times 3}$ in the (2×3) case. Although we aim at

analyzing $M > 1$ of such tables simultaneously, we abstain from further indexing whenever possible in order to increase readability. At a given genetic locus numbered by $i \in \{1,\ldots,M\}$, we want to test the null hypothesis H_0 of no association of phenotype and genetic marker i against its alternative hypothesis H_1 that phenotype and marker i are associated. We will assume that the two-sided alternative hypothesis H_1 is considered, unless stated otherwise.

In any case, the conditional probability of observing $\mathbf{x}$ given $\mathbf{n}$ under H_0 will be denoted by $f(\mathbf{x}|\mathbf{n})$ and is (in a compact, self-explaining notation) given by

$$f(\mathbf{x}|\mathbf{n}) = \frac{\prod_{n\in\mathbf{n}} n!}{N!\prod_{x\in\mathbf{x}} x!}.$$

In the remainder of this section, we review two common testing strategies for evaluating a single contingency table. A detailed survey of exact methods for contingency table analyses is provided by Agresti (1992). Moreover, we describe classical methods to control errors if many of such tables are simultaneously under consideration. The latter is important if many markers shall be tested with respect to their association with a dichotomous phenotype under the scope of one study (including meta analyses).

2.2 Marginal tests for a single contingency table

The chi-squared statistic Q for assessing association of the phenotype and the genetic marker from the observed data $\mathbf{x}$ is given by

$$Q(\mathbf{x}) = \sum_r \sum_c \frac{(x_{rc} - e_{rc})^2}{e_{rc}},$$

where r runs over the rows and c over the columns of $\mathbf{x}$ and the numbers $e_{rc} = n_{r.}n_{.c}/N$ denote the expected cell counts given N and the marginal counts contained in $\mathbf{n}$. Large values of $Q(\mathbf{x})$ are in favor of the alternative hypothesis that phenotype and genetic marker are associated.

If N is small and a confirmatory analysis with strict type I error control is required, it is not recommendable to employ the asymptotic χ^2 distribution of Q for inferential purposes, cf. Weir (1996), Wigginton et al. (2005). An exact test guaranteeing conservative type I error control is based on the p-value

$$p_Q(\mathbf{x}) = \sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}}|\mathbf{n}),$$

where the summation is carried out over all tables $\tilde{\mathbf{x}}$ with marginals $\mathbf{n}$ for which $Q(\tilde{\mathbf{x}}) \geq Q(\mathbf{x})$. For a fixed significance level α, a test φ_Q of level α is given by $\varphi_Q(\mathbf{x}) = \mathbf{1}_{p_Q(\mathbf{x})\leq\alpha}$.

An exact test of Fisher-type for testing H_0 against H_1 bases its decision directly upon $\mathbf{x}$ and utilizes a p-value

$$p_{\text{Fisher}}(\mathbf{x}) = \sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}}|\mathbf{n}),$$

where the summation is now carried out over all tables $\tilde{\mathbf{x}}$ with marginals $\mathbf{n}$ for which $f(\tilde{\mathbf{x}}|\mathbf{n}) \leq f(\mathbf{x}|\mathbf{n})$. Again, a corresponding level α test is given by $\varphi_{\text{Fisher}}(\mathbf{x}) = \mathbf{1}_{p_{\text{Fisher}}(\mathbf{x}) \leq \alpha}$.

For our approach of realized randomized p-values, presented in Section 3.2, it turns out that the chi-squared and Fisher-type testing strategies are convenient to handle.

2.3 Multiplicity correction

Let us assume a statistical model $(\Omega, \mathscr{A}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ parametrized by $\vartheta \in \Theta$. In an association study under the setup described in Section 2.1, we consider a parameter vector $\vartheta = (\vartheta_i, i = 1, \ldots, M)$. Given the marginals $\mathbf{n}_i$ for every of the M tables under consideration, the meaning and the dimensionality of the marginal parameter ϑ_i is dependent on if a (2×2)- or a (2×3)-table is considered. In the (2×2)-table case, both the genotype G_i (say) at genetic position i and the phenotype Y are binary and ϑ_i may be formalized by the probability that both G_i and Y equal zero or, equivalently, by the odds ratio (see Chapter 10 in Ziegler and König, 2006). In the (2×3)-table case, ϑ_i is two-dimensional and can be formalized by any pair of expected cell counts in the (2×3)-table corresponding to locus i, where the cells are not located in the same column. Multiple hypotheses testing is concerned with testing a family $\mathscr{H} = (H_i, i \in I)$ of hypotheses regarding the parameter ϑ with corresponding alternatives $K_i = \Theta \setminus H_i$, where I denotes an arbitrary index set. In the association study case, every genetic locus i reflects one hypothesis, namely, that G_i is stochastically independent of Y. Therefore, we simply have $I = \{1, \ldots, M\}$. For example, in the case of allelic tests in (2×2)-tables, this hypothesis translates to the parameter ϑ_i in that we test the point hypothesis that the odds ratio equals 1. Let $I_0 \equiv I_0(\vartheta) \subseteq I$ denote the index set of true hypotheses in $\mathscr{H}$, $\varphi = (\varphi_i, i \in I)$ a multiple test procedure for $\mathscr{H}$, and $V(\varphi)$ the number of false rejections of φ, i. e., $V(\varphi) = \sum_{i \in I_0} \varphi_i$. The classical multiple type I error measure is the family-wise error rate, FWER for short, and can (for a given $\vartheta \in \Theta$) be expressed as $\text{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta(V(\varphi) > 0)$. There exist various principles for constructing multiple tests controlling the FWER, meaning that $\sup_{\vartheta \in \Theta} \text{FWER}_\vartheta(\varphi) \leq \alpha$ for a pre-defined significance level α, like the intersection-union principle, the closed test principle or the partitioning principle. However, they all rely on a pre-defined

structure of $\mathscr{H}$. A universal, but often conservative method is based on the union bound and is referred to as "Bonferroni correction" in the multiple testing literature. Assuming that $|I| = M$, the Bonferroni correction carries out each individual test $\varphi_i, i \in I$, at (local) level α/M. In case that joint independence of all M marginal test statistics can be assumed, the Bonferroni-corrected level α/M can be enlarged to the "Šidák-corrected" level $1-(1-\alpha)^{1/M} > \alpha/M$ leading to slightly more powerful marginal tests. If (marginal) p-values $p_1, \ldots, p_M$ for each pair of hypotheses H_i versus K_i, $i \in I$, are available, a Bonferroni or Šidák test, respectively, controlling the FWER at level α is given by $\varphi_i = \mathbf{1}_{p_i \leq \alpha_{\text{loc.}}}$ for all $i \in I$. The local significance level $\alpha_{\text{loc.}}$ equals α/M for a Bonferroni test and $1-(1-\alpha)^{1/M}$ for a Šidák test.

Finally, we define $I_1 \equiv I_1(\vartheta) = I \setminus I_0$, $M_1 = |I_1|$, $S(\varphi) = \sum_{i \in I_1} \varphi_i$ and refer to the expected proportion of correctly detected alternatives, i. e., $\text{power}_\vartheta(\varphi) = \mathbb{E}_\vartheta[S(\varphi)/\max(M_1, 1)]$, as the multiple power of φ under ϑ. If the structure of φ is such that $\varphi_i = \mathbf{1}_{p_i \leq t^*}$ for a common, possibly data-dependent threshold t^*, then the multiple power of φ is isotone in t^*.

3 Improving the classical approaches

In Sarkar (2008a) and the subsequent discussion papers by Romano et al. (2008), Sen (2008), and Sarkar (2008b), three main challenges of modern multiple testing theory and practice are mentioned: Departure from uniform distribution of p-values under null hypotheses, appropriately taking into account dependency structures among marginal tests, and the "large M, small N" problem. We agree with this diagnosis and present some solutions under the scope of our general setup in this section.

3.1 Estimation of the proportion of informative markers

Since the index set of true hypotheses $I_0 \equiv I_0(\vartheta) \subseteq I$ depends on the unknown parameter ϑ, it is in practice not possible to control the FWER at level exactly α. The Bonferroni as well as the Šidák method bound the FWER trivially by considering I instead of I_0. In other words, these methods work under the "worst case" assumption that all M hypotheses are true. Modern (data-) adaptive multiple testing methods try to improve upon that by pre-estimating the number $M_0 = |I_0|$ or the proportion $\pi_0 = M_0/M$, respectively, of true hypotheses in $\mathscr{H}$ and replace M in $\alpha_{\text{loc.}}$ by the resulting estimation $\hat{M}_0$. It goes beyond the scope of this paper to survey all the concurring estimation techniques that are proposed in the multiple testing literature. We therefore defer the reader to the introduction in Finner and Gontscharuk (2009).

Maybe, the still most popular though, as well, the most ancient estimation technique goes back to Schweder and Spjøtvoll (1982). It relies on a tuning parameter $\lambda \in [0,1)$. Denoting the empirical cumulative distribution function (ecdf.) of the M marginal p-values by $\hat{F}_M$, the proposed estimator from Schweder and Spjøtvoll (1982) can be written as

$$\hat{\pi}_0 \equiv \hat{\pi}_0(\lambda) = \frac{1-\hat{F}_M(\lambda)}{1-\lambda}. \tag{1}$$

There exist several possible heuristic motivations for the usage of $\hat{\pi}_0$. The simplest one considers a histogram of the marginal p-values with exactly two bins, namely $[0,\lambda]$ and $(\lambda,1]$. Then, the height of the bin associated with $(\lambda,1]$ equals $\hat{\pi}_0(\lambda)$. Storey et al. (2004) and Finner and Gontscharuk (2009) investigated theoretical properties of $\hat{\pi}_0$ and slightly modified versions of this estimator. The following lemma, the proof of which is given in Appendix I, shows that $\hat{\pi}_0(\lambda)$ is a conservative estimate of π_0 with respect to its expectation. To the best of our knowledge, the bias of the Schweder-Spjøtvoll estimator has not been calculated in such generality before. Under more restrictive model assumptions (for instance, that all p-values under alternatives are stochastically independent and share the same distribution), a less general formula is given in equation (2) of Langaas et al. (2005).

Lemma 1 *The value of $\hat{\pi}_0$ is a conservative estimate of π_0, meaning that $\hat{\pi}_0$ has a non-negative bias. More specifically, it holds*

$$\mathbb{E}_\vartheta[\hat{\pi}_0(\lambda)] - \pi_0 \geq \frac{1}{M(1-\lambda)} \sum_{i \in I_1} \mathbb{P}_\vartheta(p_i > \lambda) \geq 0.$$

We will refer to this property in the discussion of Theorem 1 in Section 4.

3.2 Realized randomized p-values

The p-values defined in Section 2.2 are under null hypotheses stochastically larger than a uniformly distributed random variable on the interval $[0,1]$. This can have a massively negative impact on the multiple power of multiple testing procedures when operating with these p-values. Especially, many estimation techniques for π_0, including the Schweder-Spjøtvoll method described in the previous section, typically fail to work properly if the assumption of uniformly distributed p-values under null hypotheses is violated. This has been demonstrated in Finner et al. (2010) in the context of a discrete model with one-dimensional marginal parameters. A way out of this dilemma consists in usage of so-called "realized randomized p-values"

as defined and explained by Finner and Straßburger (2007) and Finner et al. (2010). Although they were originally derived in terms of randomized tests, we define them here in a more general way as follows.

Definition 1 *Let a statistical model $(\Omega, \mathscr{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be given. Consider the two-sided test problem $H : \{\vartheta = \vartheta_0\}$ versus $K : \{\vartheta \neq \vartheta_0\}$ and assume the decision is based on the realization $\mathbf{x}$ of a discrete random variate $\mathbf{X} \sim \mathbb{P}_\vartheta$ with values in Ω. Moreover, let U denote a uniformly distributed random variable on [0, 1], stochastically independent of $\mathbf{X}$. A realized randomized p-value for testing H versus K is a measurable mapping $p^{rand.} : \Omega \times [0,1] \to [0,1]$ fulfilling that $\mathbb{P}_{\vartheta_0}(p^{rand.}(\mathbf{X}, U) \leq t) = t$ for all $t \in [0,1]$.*

Remark 1 *It has to be mentioned at this point that randomized tests are known for a long time in the statistical literature and, for instance, build the basis for the Neyman-Pearson theory of uniformly most powerful (unbiased) tests, cf., for example, Chapter 3 in the textbook by Lehmann and Romano (2005). How to calculate p-values that are compatible with such tests is, however, a topic that is still vividly discussed in the scientific community, as the discussion of Finner and Straßburger (2007) and the recent works by Rüschendorf (2009) and Habiger and Peña (2011) show.*

The following lemma, which is a direct consequence of our more general theorem in Appendix II, provides a convenient method to compute realized randomized p-values based on the exact tests introduced in Section 2.2.

Lemma 2 *Based upon the two testing strategies described in Section 2.2, corresponding realized randomized p-values can be calculated as*

$$\begin{aligned} p_Q^{rand.}(\mathbf{x}, u) &= p_Q(\mathbf{x}) - u \sum_{\tilde{\mathbf{x}}: Q(\tilde{\mathbf{x}}) = Q(\mathbf{x})} f(\tilde{\mathbf{x}}|\mathbf{n}), \\ p_{Fisher}^{rand.}(\mathbf{x}, u) &= p_{Fisher}(\mathbf{x}) - u\kappa f(\mathbf{x}|\mathbf{n}), \end{aligned}$$

where u denotes the realization of a UNI$[0,1]$-distributed variate which is stochastically independent of $\mathbf{x}$ and $\kappa \equiv \kappa(\mathbf{x}) = |\{\tilde{\mathbf{x}} : f(\tilde{\mathbf{x}}|\mathbf{n}) = f(\mathbf{x}|\mathbf{n})\}|$.

In order to illustrate the necessity to work with realized randomized p-values in the estimation procedure described in Section 3.1, we derived Figure 1. The dashed curve in Figure 1 depicts the ecdf. of approximately $1,800$ non-randomized p-values computed from (2×3)-contingency tables with the Fisher-type testing strategy, making use of data from approximately $2,500$ randomly chosen participants in the Wellcome Trust Case Control Consortium study for the

Crohn's disease endpoint. We will provide more detail on the underlying study in Section 6.2 below. It can clearly be seen that the dashed curve partly lies below the diagonal in the unit square (which is displayed as the dotted line in Figure 1), meaning that the empirical distribution of the observed non-randomized *p*-values is stochastically larger than uniform for a non-negligible proportion of markers. For comparison, we plotted the ecdf. of the corresponding realized randomized *p*-values as the solid curve in Figure 1. After a steep increase in a neighborhood of the origin, it behaves linearly because of the defining property of realized randomized *p*-values, cf. Definition 1. Consequently, applying the estimator given in equation (1) to the *p*-values corresponding to the solid curve, we obtain a reasonable upper bound of $\hat{\pi}_0(0.5) = 0.82$ for the proportion of true null hypotheses, while the estimation procedure based on the dashed curve is almost completely uninformative and leads to $\hat{\pi}_0(0.5) = 0.9685$. Let us emphasize here that this discrepancy is not due to artifacts like, for instance, low sample size or low minor allele frequencies, but due to the inherent discreteness problem of the statistical model.

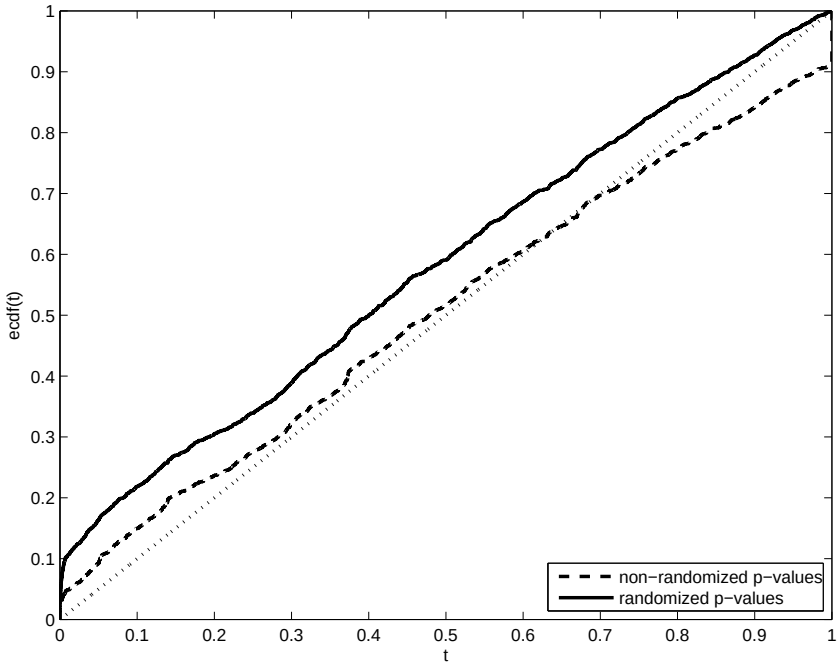

Figure 1: Empirical cumulative distribution functions of realized randomized and non-randomized *p*-values for Crohn's disease endpoint as part of The Wellcome Trust Case Control Consortium (2007) study.

3.3 Effective number of tests

Dependencies among the marginal (per marker) tests can be utilized to relax the multiplicity correction for the overall analysis. In order to motivate this heuristically, let us assume that the set of markers indexed by $I = \{1,\ldots,M\}$ can be decomposed into disjoint groups with indices in the subsets $I_g, g \in \{1,\ldots,G\}$ of I. For the moment, we now make the (unrealistic) assumption that markers within each of the subsets corresponding to the I_g's are perfectly correlated in the sense that for each $g \in \{1,\ldots,G\}$ and for any pair $(i,j) \subseteq I_g$ the identity $\{\varphi_i = 1\} = \{\varphi_j = 1\}$ holds, where $\varphi = (\varphi_1,\ldots,\varphi_M)$ is an arbitrary multiple test for the association test problem at hand. This assumption has the interpretation that in the g-th marker subgroup all tests assess the same information and therefore, "effectively" only one single test is performed in the subgroup. Denoting $i(g) = \min I_g$ for $g = 1,\ldots,G$, it is easy to check that the family-wise error rate of φ under ϑ can under the aforementioned assumptions be bounded by

$$\mathbb{P}_\vartheta\left(\bigcup_{i\in I_0}\{\varphi_i = 1\}\right) \leq \mathbb{P}_\vartheta\left(\bigcup_{g=1}^{G}\{\varphi_{i(g)} = 1\}\right),$$

with equality if every subgroup contains at least one non-associated marker. Consequently, multiplicity correction in this extreme scenario only has to be done with respect to the number G of subgroups which is typically much smaller than the number M of markers and a relaxed Bonferroni-type significance threshold for controlling the FWER is given by $\alpha/G \geq \alpha/M$.

An intuitive generalization of these simple considerations to cases in which correlation among markers is not perfect, but of arbitrary strength, is given by the Cheverud-Nyholt method for quantification of the "effective number of tests", i. e., for calculating the denominator $M_{\text{eff.}}$ in a Bonferroni-type adjustment or the exponent in a Šidák-type adjustment of the local significance levels, respectively. The formula for $M_{\text{eff.}}$ as proposed in Cheverud (2001) and Nyholt (2004) can be expressed as (see Moskvina and Schmidt, 2008)

$$M_{\text{eff.}} = 1 + \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M}(1 - r_{ij}^2). \tag{2}$$

The numbers r_{ij} in (2) are measures of correlation among markers i and j and can typically be obtained from linkage disequilibrium (LD) matrices. Linkage disequilibrium is the technical way to refer to correlations between the allelic states of different genetic markers in the same chromosome, see Lewontin and Kojima (1960). In human populations some combinations of alleles along the same chromosome (haplotypes) occur at frequencies that are different from what would be

expected out of random combinations of the markers' allelic frequencies. These correlations between markers "effectively" reduce the number of tests performed with different markers.
Despite its simplicity and intuitive character, the Cheverud-Nyholt method can not be recommended in practice, because any LD-matrix contains many values $r_{ij} = 0$ by definition of the linkage disequilibrium (marker pairs from different chromosomes have LD-coefficient equal to zero) and due to the fact that LD can only be calculated in a limited window size. These structural zeros result in very conservative (large) values of $M_{\text{eff.}}$ in practice.

A refined measure $K_{\text{eff.}}$ of the effective number of tests has been derived by Moskvina and Schmidt (2008). The authors prove that for a given LD-matrix there exists a tuple $(K_{\text{eff.}}, \alpha_{\text{loc.}})$ such that

$$\text{FWER}(\varphi) \leq 1 - (1 - \alpha_{\text{loc.}})^{K_{\text{eff.}}}, \text{ whereby } \varphi_j(\mathbf{x}_j) = \mathbf{1}_{p_Q(\mathbf{x}_j) \leq \alpha_{\text{loc.}}}. \tag{3}$$

For computation of $(K_{\text{eff.}}, \alpha_{\text{loc.}})$, they define $r_m := \max_{j=1,\ldots,m-1} |r_{jm}|$. The value r_m quantifies the largest correlation of marker $m \geq 2$ with any of the preceding markers (according to some pre-defined ordering). Now, the formula for $(K_{\text{eff.}}, \alpha_{\text{loc.}})$ is given by

$$K_{\text{eff.}} \equiv K_{\text{eff.}}(\alpha_{\text{loc.}}, (r_m)_{m \geq 2}) = 1 + \sum_{m=2}^{M} \kappa_m,$$

where κ_m depends on r_m and $\alpha_{\text{loc.}}$. An easy-to-implement numerical approximation is given by $\kappa_m = \sqrt{1 - r_m^{-1.31 \times \log_{10}(\alpha_{\text{loc.}})}}$. By means of iterative modification of $\alpha_{\text{loc.}}$ and calculation of (3), it is possible to determine $(K_{\text{eff.}}, \alpha_{\text{loc.}})$ such that the FWER is controlled at the pre-defined overall significance level α.

Although not stated explicitly, the proof by Moskvina and Schmidt (2008) only considers (2×2)-tables in connection with the chi-squared test. However, it is possible to extend their proof to the (2×3)-table case in connection with the chi-squared test. Since only the sequentially maximum LD values r_m are involved in computing $K_{\text{eff.}}$, this estimate is more appropriate in practical situations with only partially available LD information. In case that exact tests of Fisher-type are to be performed, the method of proof in Moskvina and Schmidt (2008) and, consequently, usage of $K_{\text{eff.}}$, seems not applicable. Permutation tests are a convenient alternative from the theoretical point of view, but in a genome-wide association (GWA) study with $500,000$ or one million SNPs under consideration it is often too time consuming as reported for instance in Gao et al. (2010). Therefore, we recommend permutation based estimation of the effective number of tests only for studies with a small or moderate number of candidate SNPs to be tested according to the Fisher-type testing strategy. Since the exact size of a dataset that can still be analyzed by a permutation test strategy depends on the hardware resources available

and the projected time frame for the analysis, we have to abstain from defining exact numbers for the regimes "small" and "moderate". For large M in connection with exact tests of Fisher-type, the simple $\mathscr{M}$ method derived by Gao et al. (2008) making use of a principle component analysis of the composite linkage disequilibrium (CLD) correlation matrix of the markers under consideration is recommendable. In any case, for our proposed method described in Algorithm 1 below, the correlation information (quantified as an LD or CLD matrix) has to be obtained avoiding interrelation with the association structure to be examined. We will discuss possibilities to ensure this requirement in Remark 3.

Remark 2 *It is important to notice that many artificial sources for dependencies among genetic markers exist that can not be attributed to linkage disequilibrium. In particular, absence of the Hardy-Weinberg equilibrium (HWE) in controls can induce correlations that interfere with the association analysis rather than being informative. Therefore, we assume for our analyses that a quality control procedure has been performed prior to the application of our methods and that only markers passing quality control criteria, including a test for HWE in controls, are present in the dataset at hand.*

3.4 An algorithm for improved association analyses

According to the considerations in the preceding sections, we propose the following workflow for assessing association of a binary phenotype with any of the M markers from a list of M candidates.

Algorithm 1

1. *For $j = 1,\ldots,M$, build the contingency table $\mathbf{x}_j$ carrying the information gathered for association of marker j and the phenotype under investigation.*
2. *For $j = 1,\ldots,M$, compute the realized randomized p-value $p^{rand.}(\mathbf{x}_j, u_j)$ and the non-randomized version $p(\mathbf{x}_j)$ by making use of one of the testing strategies described in Section 2.2 and the realization u_j of an UNI$[0,1]$-distributed random variable which is stochastically independent of $\mathbf{X}_j$.*
3. *Compute $\hat{\pi}_0(\lambda)$ by calculating the ecdf. of $(p^{rand.}(\mathbf{x}_j, u_j), j = 1,\ldots,M)$. In practice, it is convenient to use the value $\lambda = 0.5$ for the tuning parameter.*
4. *Determine the effective number of tests by utilizing correlation values obtained from an appropriate (C)LD matrix of the M markers. Any of the methods described before may be employed. Denote the resulting (estimated) effective number of tests by Eff.*

5. *For a pre-defined FWER level α, determine the list of associated markers by performing the multiple test $\varphi = (\varphi_j, j = 1,\ldots,M)$, where $\varphi_j(\mathbf{x}_j) = \mathbf{1}_{p(\mathbf{x}_j)\leq t^*}$ with $t^* = \alpha/(Eff\cdot\hat{\pi}_0(\lambda))$.*

Remark 3

(a) Notice that we propose to use the computed realized randomized p-values in the third step of Algorithm 1 while for final decision making in step 5 the non-randomized p-values are to be used. This policy ensures accurate estimation of π_0 on the one hand and reproducibility of the test result on the other hand. It may be argued that the estimated value of π_0 also depends on the realization of the uniform variates used for randomization. But, first of all, as demonstrated by Finner et al. (2010), the variance of $\hat{\pi}_0$ with respect to the distribution of these uniform variates is typically very small. Secondly, it is possible to replace the value of $\hat{\pi}_0$ by its conditional expectation with respect to randomization, computed in Appendix III.

(b) The underlying assumption of Algorithm 1 is that the pairwise marker correlations are on average of not smaller magnitude in the group of markers which are not associated with the phenotype under investigation than in the group of informative markers. This assumption can be formalized as the relationship

$$\pi_0 = \frac{M_0}{M} \geq \frac{Eff(I_0)}{Eff} \quad \textit{or, equivalently,} \quad \pi_0 Eff \geq Eff(I_0), \tag{4}$$

where $Eff(I_0)$ denotes the effective number of tests within the subset of markers for which the null hypothesis of no association with the phenotype holds. Of course, assumption (4) cannot be verified in practice, because I_0 is unobservable. However, it seems very natural to us, because informative markers are assumed to be sparsely distributed among the genome and consequently most of their pairwise LD values should be of low magnitude. Non-associated markers (with the phenotype), however, lie dense and should have on average a higher pairwise correlation.

Moreover, two natural possibilities exist to ensure that experimental conditions cannot lead to a confounding influence of the disease status on the LD values utilized to calculate Eff (Such a confounding influence might lead to violations of (4)*.):*

1. *Assessing (C)LD information from an external reference database instead of estimating it from the actual data sample under investigation*
2. *Performing computation of Eff only in the subgroup of control individuals*

In practice, the second method seems more convenient and is a quasi-standard technique in the genetics community. Even if reference samples like the HapMap database are available, it is not guaranteed that they are perfectly representative for the data sample in a particular study.

The following theorem shows that in large-scale investigations (like in a genome-wide scan) the FWER is controlled at level α by Algorithm 1 if we use the Moskvina and Schmidt (2008) method, which is favored by us.

Theorem 1 *Let assumption (4) be fulfilled and let the effective number of tests be estimated by $K_{eff.}$ according to Moskvina and Schmidt (2008) for the chi-square testing strategies. If the cumulative distribution function (cdf.) of $(p^{rand.}(\mathbf{x}_j, u_j), j \in I_0)$ converges to the cdf. of UNI$[0,1]$ for $M_0 \to \infty$, Algorithm 1 asymptotically ($M_0 \to \infty$) controls the FWER at level α.*

Proof: The estimate $K_{\text{eff.}}$ is deduced by probabilistic upper bounds guaranteeing that it is an upper bound itself in the sense that the FWER is strictly controlled at level α if the threshold $1-(1-\alpha)^{1/K_{\text{eff.}}}$ for the p-values is used, even if all M null hypotheses are true. Furthermore, Lemma 2 in Finner and Gontscharuk (2009) shows that convergence of the cdf. of p-values corresponding to non-informative markers to the cdf. of UNI$[0,1]$ implies that $\hat{\pi}_0$ estimates π_0 asymptotically almost surely conservatively in the sense that $\liminf_{M_0\to\infty}\{\hat{\pi}_0/\pi_0\} \geq 1$ $[\mathbb{P}_\vartheta]$ for all possible parameter values ϑ of the statistical model. The assertion now follows by noticing that $\alpha/\ell < 1-(1-\alpha)^{1/\ell}$ for all $\ell \geq 1$. ■

In case of the Fisher-type testing methods and usage of simple $\mathscr{M}$, an analogous result can be obtained if the tuning parameter C of simple $\mathscr{M}$ is chosen conservatively, cf. Gao et al. (2008).

4 Small-scale simulation study

The assertion of Theorem 1 is an asymptotic one for the number M of markers under investigation tending to infinity. As far as exact finite FWER control is concerned, we return to the assertion of Lemma 1. We have shown that the Schweder-Spjøtvoll estimator $\hat{\pi}_0$ estimates π_0 conservatively with respect to its first moment. In other words, on average we expect an overestimation of π_0 by $\hat{\pi}_0$. Now, the investigations in Section 2 of Finner and Gontscharuk (2009) show that slightly modified versions of $\hat{\pi}_0$ provide exact finite FWER control if the p-values under null hypotheses are stochastically independent. The authors propose to add a constant in its nominator making the estimator more conservative than just with respect to its first moment.

For arbitrarily dependent *p*-values, the situation is more complicated. However, in the association analysis case with two-sided alternatives as considered here, only positive dependency can occur. As recently studied extensively in the context of false discovery rate (FDR) theory (cf., e. g., Benjamini and Yekutieli (2001), Sarkar (2002), Finner et al., 2007), multiple tests typically behave more conservatively under positive dependency than under joint independence.

Anyhow, in order to assess the behavior of Algorithm 1 for small numbers of markers under investigation (as, for instance, in replication studies), we performed a small-scale simulation study for different (small and moderate) values of M, and with $M_1 = 10$ in all cases. This parameter setup has been chosen to roughly reflect the situation in Section 6.1 below. To this end, since simulation of real synthetic genetic data is a very complicated task, we made use of *semi-synthetic* data, meaning that we used true observed genotypes (taken from the WTCCC Crohn's disease sub-study which we will describe in Section 6.2 below), and only brought the disease indicators under experimental control. More specifically, we employed a logistic regression model with additive risk allele contributions of the form

$$\mathbb{P}_\beta(Y_i = 1|G_i) = [1+\exp(-z_i)]^{-1}, \text{ where } \quad z_i = \gamma \sum_{j=1}^{M_1} \beta_j \cdot G_{i,j}. \tag{5}$$

In equation (5), the index $1 \leq i \leq N = 4,688$ corresponds to individuals and the index $1 \leq j \leq M_1$ runs over the M_1 positions on the genome which have been chosen to carry information about the phenotype.

In order to ensure that the M_0 positions chosen to be uninformative in these computer simulations can really fulfill this requirement, they must be uncorrelated with the M_1 positions chosen to contain the "signals". Therefore, for practical implementation, we implanted a subsequent block (in terms of the ordering of the genetic positions present in the raw data) of M_1 markers from chromosome 2 into a subsequent block of M_0 markers from chromosome 1. This ensures a realistic LD structure within both blocks. The blocks were chosen randomly, but we ensured a minor allele frequency of at least 10% at every locus included in the simulation data sets.

For ease of exposition and since these simulations shall mainly serve as a proof of principle, the regression coefficients $\beta = (\beta_1, \ldots \beta_{M_1})^t$ have been drawn independently and uniformly from the interval $[-1,1]$ and normalized such that they summed up to 0. Neither the LD structure nor the proportion π_0 is affected by the choice of β, as long as all coefficients β_j are different from zero. The genotype

information $G_{i,j}$ for individual i at locus j was coded as follows.

$$\begin{aligned} G_{i,j} &= 0, \quad \text{if the genotype of individual } i \text{ at locus } j \text{ equals } A_1A_1, \\ G_{i,j} &= 1, \quad \text{if the genotype of individual } i \text{ at locus } j \text{ equals } A_1A_2, \\ G_{i,j} &= 2, \quad \text{if the genotype of individual } i \text{ at locus } j \text{ equals } A_2A_2, \end{aligned}$$

with A_1 denoting the wild type and A_2 denoting the risk allele (variant) at locus j. The "attenuation factor" γ in equation (5) should reflect the fact that other covariates apart from the genotype have an influence on the phenotype, too. We chose $\gamma = 1/4$ in our simulations, leading to realistic effect sizes in terms of the empirical distribution of the p-values corresponding to the M_1 informative positions (compared with the reported p-values entailing strong and moderate evidence for association in the WTCCC Crohn's disease sub-study).

For every setup (every considered value of M), we performed $B = 1000$ Monte Carlo repetitions of the following simulation algorithm.

Algorithm 2

1. *Draw disease labels according to the model in equation* (5)*.*
2. *Apply (a) the Bonferroni correction, (b) the Bonferroni plug-in method from Finner and Gontscharuk (2009), (c) the method from Moskvina and Schmidt (2008), (d) Algorithm 1, to the simulated data.*
3. *Record for all four methods (a) if a type I error occurred, (b) the number of truly associated positions (with the phenotype) that could be detected.*

After completion of all $B = 1000$ Monte Carlo iterations, we estimated the family-wise error rate and the multiple power of the four concurring multiple tests by relative frequencies and means, respectively. The results are summarized in Table 1.

For $M = 50$, all three data-adaptive methods behaved liberally, with Algorithm 1 showing the largest empirical exceedance of the nominal FWER level. Similarly as in Section 2 of Finner and Gontscharuk (2009), one can calibrate either the nominator of the estimator $\hat{\pi}_0$ or the nominal α to be utilized in Algorithm 1 for very small numbers of M such that exact control of the FWER is ensured, if a concrete genetic model can be assumed. If the latter is not the case, a simple ad-hoc adjustment of α can be based on computer simulations of the type described in the present section and by noticing that t^* is a linear function of α. However, it has to be warned that this type of adjustment does not imply a strict (mathematically proven) guarantee for FWER control. This is a drawback of all data-adaptive procedures that implicitly rely on asymptotic theory like the Glivenko-Cantelli theorem.

	$M = 50, M_0 = 40,$ $\overline{\hat{\pi}_0} = 0.859, K_{\text{eff.}} = 38.21$	$M = 60, M_0 = 50,$ $\overline{\hat{\pi}_0} = 0.8764, K_{\text{eff.}} = 45.65$
$\widehat{\text{FWER}}$(Bonf.)	0.040	0.040
$\widehat{\text{FWER}}$(BPI)	0.053	0.050
$\widehat{\text{FWER}}$(MS)	0.053	0.050
$\widehat{\text{FWER}}$(Alg. 1)	0.063	0.063
$\widehat{\text{power}}$(Bonf.)	0.2932	0.1791
$\widehat{\text{power}}$(BPI)	0.3068	0.1894
$\widehat{\text{power}}$(MS)	0.3184	0.1983
$\widehat{\text{power}}$(Alg. 1)	0.3343	0.2100
	$M = 65, M_0 = 55,$ $\overline{\hat{\pi}_0} = 0.9052, K_{\text{eff.}} = 49.67$	$M = 70, M_0 = 60,$ $\overline{\hat{\pi}_0} = 0.9105, K_{\text{eff.}} = 52.35$
$\widehat{\text{FWER}}$(Bonf.)	0.034	0.026
$\widehat{\text{FWER}}$(BPI)	0.041	0.031
$\widehat{\text{FWER}}$(MS)	0.043	0.036
$\widehat{\text{FWER}}$(Alg. 1)	0.054	0.045
$\widehat{\text{power}}$(Bonf.)	0.2486	0.1699
$\widehat{\text{power}}$(BPI)	0.2560	0.1763
$\widehat{\text{power}}$(MS)	0.2652	0.1877
$\widehat{\text{power}}$(Alg. 1)	0.2725	0.1974
	$M = 75, M_0 = 65,$ $\overline{\hat{\pi}_0} = 0.9161, K_{\text{eff.}} = 56.45$	$M = 100, M_0 = 90,$ $\overline{\hat{\pi}_0} = 0.9405, K_{\text{eff.}} = 75.58$
$\widehat{\text{FWER}}$(Bonf.)	0.035	0.033
$\widehat{\text{FWER}}$(BPI)	0.039	0.036
$\widehat{\text{FWER}}$(MS)	0.040	0.042
$\widehat{\text{FWER}}$(Alg. 1)	0.048	0.047
$\widehat{\text{power}}$(Bonf.)	0.7085	0.7054
$\widehat{\text{power}}$(BPI)	0.7092	0.7055
$\widehat{\text{power}}$(MS)	0.7132	0.7081
$\widehat{\text{power}}$(Alg. 1)	0.7141	0.7089

Table 1: Simulation results on semi-synthetic data. Abbreviation ”Bonf.” refers to the Bonferroni correction, ”BPI” to Bonferroni plug-in, ”MS” to the Moskvina and Schmidt (2008) method, and ”Alg. 1” to Algorithm 1. The target FWER level was set to $\alpha = 5\%$ in all simulations.

For $M = 60$ and $M = 65$, the liberal behavior of Algorithm 1 was still observed (with decreasing severity), but already for $M = 70$, it did not occur anymore and Algorithm 1 exhausted the nominal FWER level best among all four methods for $M = 70$, $M = 75$, and $M = 100$. If one would change the distribution of the vector β of regression coefficients such that most p-values corresponding to alternatives are close to the decision boundary t^*, one could construct situations in which exhaustion of the FWER level also translates in a more pronounced way into gain in multiple power.

Remark 4 *All simulations in this section have been run on a standard quad-core desktop personal computer. For one simulation setup (one value of M), they took between* 8 *and* 8.5 *hours (drawing of* $1000 \times N \approx 4,700$ *labels, computation of* $1000 \times M$ *non-randomized and realized randomized p-values, estimation of* π_0 1000 *times, computation of* $K_{eff.}$*, final evaluation with respect to FWER control and multiple power). For carrying out the computations for a genome-wide analysis as described in Section 6.2, we recommend to make use of cluster-computing techniques such that computations can be parallelized, for instance with respect to chromosomes. Computing time in this case will depend on many factors such as general workload of the cluster, availability of physical and virtual memory, etc. As far as software and programming is concerned, we provide hints for efficient implementation in the next section.*

5 Computational details

The main computational complexity of the algorithm described in Section 3.4 originates from the necessity to traverse all tables $\tilde{\mathbf{x}}$ with given marginals $\mathbf{n}$ in order to compute realized randomized p-values in the second step of Algorithm 1, because the ordering induced by $Q(\cdot)$ or $f(\cdot|\mathbf{n})$, respectively, cannot be utilized in a straightforward way, meaning that it is hardly possible to determine the set of $\tilde{\mathbf{x}}$'s to be summed over explicitly.

To derive a feasible implementation, we first notice that the logarithmic conditional probability of observing $\mathbf{x}$ given $\mathbf{n}$ can be expressed as $\ln(f(\mathbf{x}|\mathbf{n})) = A(\mathbf{n}) - B(\mathbf{x})$ with $A(\mathbf{n}) = \sum_{n \in \mathbf{n}} \ln(\Gamma(n+1)) - \ln(\Gamma(N+1))$ and $B(\mathbf{x}) = \sum_{x \in \mathbf{x}} \ln(\Gamma(x+1))$. Thereby, $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$ denotes the Gamma function. This decomposition in a term only depending on $\mathbf{n}$ and another term only depending on $\mathbf{x}$ is extremely helpful, because $A(\mathbf{n})$ can be pre-computed before iterating over the $\tilde{\mathbf{x}}$'s. Moreover, the transformation with the natural logarithm stabilizes computations and protects against integer overflow. The additive structure of $\ln(f(\mathbf{x}|\mathbf{n}))$ has the

additional merit that it can be evaluated very efficiently by computer software, especially MATLAB, which provides the fully vectorized function `gammaln` for evaluating the logarithmic Gamma function.

For implementing the iterations over the possible tables $\tilde{\mathbf{x}}$, it is essential to notice that, given $\mathbf{n}$, each (2×2) table $\mathbf{x}$ is already uniquely defined by the entry x_{11}. All other entries of $\mathbf{x}$ can be calculated from x_{11} and $\mathbf{n}$. This means, a single loop over the possible values of x_{11} suffices to traverse all tables. In the (2×3) table situation, a double loop over x_{11} and x_{12} is sufficient. Furthermore, one can restrict the number of tables to be traversed even further by incorporating all constraints on the entries of the $\tilde{\mathbf{x}}$'s given by the marginals $\mathbf{n}$. More specifically, in the (2×2) table situation, x_{11} has to be a member of the set $\{\max(0, n_{1\cdot} - n_{\cdot 2}), \ldots, \min(n_{1\cdot}, n_{\cdot 1})\}$. In case of a (2×3) table, it necessarily holds $x_{11} \in \{0, \ldots, \min(n_{1\cdot}, n_{\cdot 1})\}$ and (as soon as the value of x_{11} is fixed) $x_{12} \in \{\max(0, n_{1\cdot} - n_{\cdot 2} - x_{11}), \ldots, \min(n_{1\cdot} - x_{11}, n_{\cdot 1})\}$.

As supplementary material, we provide four efficient MATLAB routines for calculating (non-)randomized p-values p_Q and p_{Fisher} for both (2×2) and (2×3) tables upon request. We like to acknowledge Giuseppe Cardillo's implementation `myfisher23`, cf. Cardillo (2007), which already features many of the aforementioned implementational tricks except some restrictions on x_{11} and x_{12} and the computation of realized randomized p-values. Furthermore, corresponding `R` routines will be included in the next release of the μTOSS software system for multiple comparisons by Blanchard et al. (2010).

6 Performance on real-life datasets

6.1 Replication study by Herder et al. (2008), type II diabetes endpoint

The study reported by Herder et al. (2008) aimed at replicating genetic variants conferring an increased type II diabetes risk in a population in Southern Germany. To this end, $M = 44$ SNPs on ten different genes were considered. In the "Results" section, the authors state that a "(conservative) Bonferroni correction for 10 genes" leads to a FWER-controlling multiple test procedure for this dataset. Setting the FWER level to $\alpha = 5\%$, this correction means that a threshold of 0.005 has to be used for raw marginal p-values. However, the claimed conservativeness is only guaranteed in the artificial situation we discussed at the beginning of Section 3.3, i. e., if all markers within a gene are perfectly correlated ($r_{ij}^2 = 1$). We re-analyzed the data according to Algorithm 1. Before discussing the results it is worth mentioning that the original study performed allelic (odds ratio-based) tests with simultaneous

adjustment for covariates gender, age and body-mass index. Since a deep discussion about the validity of Algorithm 1 in case of adjustment for covariates is way beyond the scope of our work, we abstained from adjusting for covariates and only analyzed the genetic component of the associations. However, as shown in Table 2, adjustment for covariates changes p-values and odds ratios only marginally so that it seems justified not to consider adjustments here. For shortness of presentation, we only include the 13 SNPs on chromosomes 3 and 6 in Table 2; the results for the remaining 31 SNPs are very similar and can be found in Appendix IV.

SNP	Allelic OR (adjusted)	one-sided p (adjusted)	Allelic OR (unadjusted)	one-sided $p^{\text{rand.}}$ (unadjusted)
rs11709077	0.74	0.0078	0.7668	0.0114
rs17036328	0.77	0.015	0.7911	0.0235
rs1801282	0.76	0.010	0.7764	0.0144
rs16860234	1.12	0.11	1.1357	0.0791
rs4402960	1.11	0.11	1.1258	0.0792
rs7651090	1.10	0.13	1.1111	0.1075
rs7640744	1.07	0.23	1.0806	0.1850
rs1470579	1.15	0.0499	1.1634	0.0403
rs10946398	1.30	0.00084	1.2661	0.0019
rs7754840	1.30	0.00073	1.2695	0.0017
rs9460546	1.30	0.00075	1.2695	0.0021
rs9465871	1.39	0.00040	1.3343	0.0015
rs7767391	1.37	0.00059	1.3164	0.0020

Table 2: Odds ratios and p-values for the first real data example

Utilizing LD information from HapMap (population 'CEU'), we applied the Moskvina-Schmidt method for computing the effective number of tests and obtained $K_{\text{eff.}} = 16.73$. As expected (different chromosomes involved in the analysis), the Cheverud-Nyholt method leads to a very conservative estimation of $M_{\text{eff.}} = 40.63$. Notice that incorporating the effective number of tests alone does not reduce multiplicity to the "Bonferroni regarding number of genes"-type threshold mentioned before. However, additional estimation of the proportion of uninformative markers leads to $\hat{\pi}_0 = 0.4545$ and, altogether, Algorithm 1 leads to the threshold $t^* = 0.0066$ for the raw p-values. Even if we would calibrate the nominal FWER level to be employed in Algorithm 1 in such a way that our simulations indicate that the target FWER level of $\alpha = 5\%$ is strictly kept for this small number of $M = 44$, the corresponding new threshold would still exceed 0.005. In summary, our proposed method confirms the heuristic argumentation in Herder et al. (2008) and endorses that the *CDKAL1* gene has been replicated in their study.

Remark 5 *For the estimation of π_0 in case of one-sided p-values, we utilized the slightly modified technique from Barras et al. (2010).*

6.2 WTCCC dataset, Crohn's disease endpoint

Here, we demonstrate the usefulness of our new method for the case of a genome-wide association analysis. To this end, we re-analyzed the dataset for Crohn's disease as part of The Wellcome Trust Case Control Consortium (WTCCC) study, cf. The Wellcome Trust Case Control Consortium (2007), consisting of $455,086$ SNPs and $4,688$ individuals (after quality control). Our proposed workflow in the GWA case consists of two stages, a screening and a validation stage, as already considered by Evans et al. (2009), for example. To this end, we performed the following procedure on the data.

(i) Split the WTCCC Crohn's disease sample randomly into two halves, but keeping the ratio cases / controls constant in both subsamples.
(ii) Consider the first sub-sample and apply an FDR-controlling (screening) criterion to generate a list of candidate SNPs (there will be false positives in this list).
(iii) Apply Algorithm 1 to the second subsample, but only considering the detected candidate SNPs from the first subsample.

Our analysis can be regarded as a confirmatory pseudo-experiment consisting of the two stages mentioned before. Of course, if all data are available for a combined analysis, we do *not* recommend to split it. The aforementioned procedure shall only mimic our target situation where a two-stage data ascertainment design has been planned beforehand in order to pre-screen a set of candidate markers. Such a data analysis strategy is often chosen in practice. In such a design, it would statistically not be valid to combine the data for the pre-screened markers for final analysis. The reasons that we used the data from the WTCCC study for this illustrative purpose are that these data are well-known, of validated high quality and consisting of many individuals.

In step (ii) of our analysis, we set the FDR level to $q = 1/2$, meaning that we expect half of the output positions truly non-associated, but also ensuring that most of the truly associated positions should be present in the output list. Indeed, application of the FDR criterion with this parameter led to an adjusted threshold of 0.0026 for realized randomized *p*-values from the first sub-sample and selected (almost) a superset of size $1,778$ of the SNPs reported as associated with Crohn's disease in Tables 3 and 4 of The Wellcome Trust Case Control Consortium (2007),

as expected, although we have drastically reduced power in comparison with utilizing the full dataset. Only one position on chromosome 19 that appears in Table 4 of The Wellcome Trust Case Control Consortium (2007) could not be detected using the FDR criterion.

In step (iii), we made use of LD coefficients computed from all controls (which is valid, because there is no interrelation with the phenotype). For assessing the stability of $K_{\text{eff.}}$, we first performed computation of the effective number of tests twice in the entire sample, once for LD computed in a window size of 10 kilobases and once for a 100 kilobases window. Usage of a 10 kilobases window resulted in an estimated effective number of tests of $K_{\text{eff.}} = 346,167.96$ and utilizing the more informative LD-values in the 100 kilobase window led to $K_{\text{eff.}} = 329,079.66$. For determining the final threshold for the $1,778$ p-values corresponding to the positions selected in step (ii) and computed from the second sub-sample in step (iii), we used the 100 kilobase window and obtained $K_{\text{eff.}} = 1,350.45$. Additionally, we computed $\hat{\pi}_0(1/2) = 0.820$ as already mentioned in the discussion of Figure 1 and arrived at a multiplicity-adjusted threshold $t^* = 4.515 \times 10^{-5}$ for p-values originating from the second sub-sample (the FWER level was set to $\alpha = 0.05$ as in the original publication).

As shown in Table 3, the final output dataset consists of 24 genetic positions that could be detected to have a significant association with Crohn's disease and is in good concordance with the results obtained by The Wellcome Trust Case Control Consortium (2007).

7 Discussion

First, we discuss briefly how to choose between the two concurring marginal testing strategies described in Section 2.2. Common knowledge among statisticians and practitioners seems to be, on the one hand, that even for larger sample sizes, exact tests of Fisher-type tend to behave conservatively. On the other hand, chi-squared tests are considered inappropriate for small sample sizes, because they originate from asymptotic considerations and because the chi-squared statistic Q is very sensitive with respect to small expected cell counts e_{rc} in its denominator. However, these two general properties of the Fisher and the chi-squared tests are of qualitative character and they do not yet allow for the choice of a testing strategy for a concrete dataset at hand. A quantitative assessment of the degree of conservativeness of Fisher's exact test can be found in Crans and Shuster (2008). The authors also provide a numerical remedy by tabulating adjustment constants leading to an exhaustion of the (marginal) significance level by Fisher's exact test. Lydersen et al. (2009) provide a biostatistics tutorial with practical guidelines for choosing

Chromosome	SNP	two-sided p-value
1	rs11805303	1.52815×10^{-6}
1	rs10489629	4.32475×10^{-5}
1	rs2201841	8.98027×10^{-6}
2	rs10210302	1.37414×10^{-8}
2	rs6752107	1.24311×10^{-8}
2	rs6431654	1.46089×10^{-8}
2	rs3828309	3.70906×10^{-8}
2	rs3792106	1.26383×10^{-8}
5	rs17234657	9.39656×10^{-7}
5	rs9292777	7.45977×10^{-6}
5	rs1505992	5.33684×10^{-6}
5	rs1553576	2.2623×10^{-5}
5	rs1553577	1.5899×10^{-5}
5	rs4957313	2.6807×10^{-5}
5	rs6896604	3.29288×10^{-5}
5	rs4957317	2.63496×10^{-5}
5	rs11750156	6.87614×10^{-6}
5	rs10055860	8.0347×10^{-6}
5	rs1122433	7.61894×10^{-6}
5	rs11957134	3.62912×10^{-5}
5	rs1000113	4.20982×10^{-5}
5	rs11747270	1.61028×10^{-5}
10	rs11816049	3.79493×10^{-5}
16	rs2076756	3.83491×10^{-6}

Table 3: Output dataset for the second real data example

a marginal testing strategy. Our approach for working with realized randomized p-values can be regarded as a generalization of the concept of "mid p-values" proposed in the latter article.

Second, a question of practical interest is: how much gain can be expected by applying Algorithm 1 in comparison with estimation of the effective number of tests alone, for example? For instance, it may be argued that for large-scale GWA studies in which only a tiny proportion of SNPs are expected to be associated with the phenotype, the multiplicity reduction proposed in Section 3.4 will mainly be due to the incorporation of the effective number of tests and that the additional estimation of the proportion of true null hypotheses will only yield a negligible additional contribution. Although this is true, an association analysis for a yet completely unexplored phenotype typically consists of two stages: a screening and a validation stage (meant here to be carried out under the scope of one study). Our workflow proposes utilizing the same LD information (obtained from the control samples or from an external reference database) in both stages. This is especially useful if the validation data set is of much smaller sample size, making correlation estimates less stable than in the first analysis phase. The latter contradicts the notion that the second phase will provide more reliable statistical evidence. Moreover, $\hat{\pi}_0(\lambda)$ will typically be small in the validation phase, giving rise to a notable increase in multiple power in comparison to mere determination of the effective number of tests. This has been demonstrated by re-analyzing a replication study in Section 6.1 and the WTCCC data for Crohn's disease in Section 6.2.

A further point worth discussing may be the question whether FDR control is more appropriate than FWER control for genetic studies and how our methodology relates to FDR control. Finner et al. (2010) describe the usage of realized randomized p-values and $\hat{\pi}_0(\lambda)$ in connection with an FDR-based analysis for Hardy-Weinberg equilibrium. In such a case, type II error control (not to include too many markers with a lack of genotyping quality in the analysis) is of much higher importance than in the association test situation considered here, especially if no independent replication study is possible or desired. The FWER thus seems the more natural criterion in our setup. However, utilizing realized randomized p-values also in the fifth step of Algorithm 1 in the screening stage of a study might be appropriate if a following validation stage is planned beforehand. This is due to the fact that the final decisions are only made in this second (validation) stage. From a methodological point of view, the open question with respect to a possible transfer of our considerations to FDR-controlling multiple test procedures is how the effective number of tests can be incorporated appropriately in the classical linear step-up (LSU) test by Benjamini and Hochberg, for example. As shown in Finner et al. (2007), positive correlations of medium magnitude lead to a very conservative behavior of the LSU test, and a natural consequence of our work seems to adjust

the FDR level q by a factor depending on the effective number of tests and the proportion of true nulls. However, the FDR is defined as the expected value of the ratio of two dependent random variables and its value is therefore not necessarily increasing in the value of the nominator. This imposes technical problems which have not yet been resolved. Storey et al. (2004) have introduced an adjustment only making use of $\hat{\pi}_0(\lambda)$, but additional consideration of the effective number of tests has to our knowledge not been treated in the literature yet.

Finally, it will be interesting to explore how adjustment for covariates such as age, gender or socio-economic variables can influence the correlation structure assessment. This topic was briefly raised in Section 6.1, but goes beyond the scope of our work and is devoted to future research.

Appendix I: Bias of the Schweder-Spjøtvoll estimator

In order to compute the bias of $\hat{\pi}_0(\lambda)$, we have to calculate

$$\mathbb{E}_\vartheta[\hat{\pi}_0(\lambda)] = (1-\lambda)^{-1}(1-\mathbb{E}_\vartheta[\hat{F}_M(\lambda)]). \tag{6}$$

To this end, we decompose

$$\mathbb{E}_\vartheta[\hat{F}_M(\lambda)] = M^{-1}\left(\sum_{i\in I_0}\mathbb{P}_\vartheta(p_i\leq\lambda)+\sum_{i\in I_1}\mathbb{P}_\vartheta(p_i\leq\lambda)\right).$$

Due to the defining property of a p-value, i. e., $\mathbb{P}_\vartheta(p_i\leq\lambda)\leq\lambda$ for all $i\in I_0$, it holds $\mathbb{E}_\vartheta[\hat{F}_M(\lambda)]\leq\pi_0\lambda+M^{-1}\sum_{i\in I_1}\mathbb{P}_\vartheta(p_i\leq\lambda)$. Abbreviating

$$S^{\leq}=M^{-1}\sum_{i\in I_1}\mathbb{P}_\vartheta(p_i\leq\lambda)\ \text{ and }\ S^{>}=M^{-1}\sum_{i\in I_1}\mathbb{P}_\vartheta(p_i>\lambda),$$

leading to $S^{\leq}+S^{>}=1-\pi_0$, we immediately obtain that $\mathbb{E}_\vartheta[\hat{\pi}_0(\lambda)]\geq(1-\lambda)^{-1}\times(S^{>}+\pi_0(1-\lambda))$ by substituting $1=S^{\leq}+S^{>}+\pi_0$ in the second factor of (6). Thus, the bias of $\hat{\pi}_0(\lambda)$ is lower-bounded by

$$\mathbb{E}_\vartheta[\hat{\pi}_0(\lambda)]-\pi_0\geq\frac{S^{>}}{1-\lambda}=\frac{1}{M(1-\lambda)}\sum_{i\in I_1}\mathbb{P}_\vartheta(p_i>\lambda)\geq 0, \tag{7}$$

whereby the first inequality in (7) is an equality if p-values under null hypotheses are uniformly distributed on $[0,1]$.

Remark 6 *It may be worth to study the extremes of the bias of $\hat{\pi}_0(\lambda)$ under the assumption that p-values under null hypotheses are uniformly distributed on* $[0,1]$.

For the Dirac-case $p_i \sim \delta_0$ for all $i \in I_1$, we obtain that $\hat{\pi}_0(\lambda)$ is unbiased for any $\lambda \in [0,1)$ which is in line with the findings in Finner and Gontscharuk (2009). On the other hand, if $p_i \sim UNI([0,1])$ for all $i \in I$, the bias of $\hat{\pi}_0(\lambda)$ equals $(1-\pi_0)$ for any $\lambda \in [0,1)$ meaning that $\mathbb{E}_\vartheta[\hat{\pi}_0(\lambda)] = 1$.

Appendix II: Realized randomized p-values

Theorem 2 *Let $G : \Omega \to \mathbb{R}$ and let $f : \Omega \to \mathbb{R}_+$ be a density on Ω of a discrete random variate $\mathbf{X}$, such that $f(x) > 0$ for all $x \in \Omega$. Moreover let U denote a $UNI[0,1]$-distributed variate which is stochastically independent of $\mathbf{X}$. Define*

$$\begin{aligned} p_G(x) &= \sum_{y:G(y)\le G(x)} f(y), \\ p_G^{rand.}(x,u) &= \sum_{y:G(y)\le G(x)} f(y) - u \sum_{y:G(y)=G(x)} f(y), \text{ and} \\ \mathscr{W} &= \{p_G(x) : x \in \Omega\}, \end{aligned}$$

then it holds

$$\mathbb{P}(p_G(\mathbf{X}) \le t) \le t, \text{ for all } t \in [0,1], \tag{8}$$

$$\mathbb{P}(p_G(\mathbf{X}) \le t) = t, \text{ for all } t \in \mathscr{W}, \tag{9}$$

$$\mathbb{P}(p_G^{rand.}(\mathbf{X},U)) \le t) = t, \text{ for all } t \in [0,1]. \tag{10}$$

Proof: Inequality (8) follows directly from (9). To prove (9) let $t \in \mathscr{W}$. Then there exists a $z \in \Omega$ such that $t = p_G(z)$ and $p_G(\mathbf{x}) \le t$ is equivalent to $G(x) \le G(z)$ and thus,

$$\mathbb{P}(p_G(\mathbf{X}) \le t) = \mathbb{P}(G(\mathbf{X}) \le G(z)) = p_G(z) = t.$$

Similarly, one can prove (10). Note that for each $t \in [0,1]$ there exists a $q \in [0,1]$ and a $z \in \Omega$ such that $t = p_G^{\text{rand.}}(z,q)$. Now $p_G^{\text{rand.}}(x,u) \le p_G^{\text{rand.}}(z,q)$ holds if either $G(x) < G(z)$ or $G(x) = G(z)$ and $u \ge q$ holds, and we have

$$\begin{aligned} \mathbb{P}(p_G^{\text{rand.}}(\mathbf{X},U) \le t) &= \mathbb{P}(G(\mathbf{X}) < G(z)) + \mathbb{P}(G(\mathbf{X}) = G(z), U \ge q) \\ &= p_G^{\text{rand.}}(z,q) = t. \ \blacksquare \end{aligned}$$

Appendix III: Conditional expectation of $\hat{\pi}_0$

Recall that $U = (U_1, \ldots, U_M)^t$ is a vector of stochastically independent, identically uniformly on $[0,1]$ distributed random variables. Moreover, U is stochastically independent of the vector $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)^t$ of all contingency table data. Now, we consider

$$\mathbb{E}^U[\hat{\pi}_0(\lambda)|\mathbf{X} = \mathbf{x}] = \frac{1 - \mathbb{E}^U[\hat{F}_M(\lambda)|\mathbf{X} = \mathbf{x}]}{1-\lambda},$$

where $\mathbb{E}^U[\cdot]$ refers to the mathematical expectation with respect to the (joint) distribution of U and $\hat{F}_M$ denotes the ecdf. of the realized randomized p-values. Since for every $1 \leq j \leq M$ the variable U_j that is used for randomization is stochastically independent of the table data $\mathbf{X}_j$, we immediately obtain that

$$\mathbb{E}^U[\hat{F}_M(\lambda)|\mathbf{X} = \mathbf{x}] = M^{-1} \sum_{j=1}^{M} \mathbb{P}^{U_j}(p^{\text{rand.}}(\mathbf{x}_j, U_j) \leq \lambda).$$

Let $A_j = \{\tilde{\mathbf{x}} : f(\tilde{\mathbf{x}}|\mathbf{n}) = f(\mathbf{x}_j|\mathbf{n})\}$ or $A_j = \{\tilde{\mathbf{x}} : Q(\tilde{\mathbf{x}}) = Q(\mathbf{x}_j)\}$, respectively. We have to distinguish three cases: First, if the non-randomized p-value $p(\mathbf{x}_j)$ already fulfills $p(\mathbf{x}_j) \leq \lambda$, we have $\mathbb{P}^{U_j}(p^{\text{rand.}}(\mathbf{x}_j, U_j) \leq \lambda) = 1$. Second, if $p(\mathbf{x}_j) > \lambda + \sum_{\tilde{\mathbf{x}} \in A_j} f(\tilde{\mathbf{x}}|\mathbf{n})$, it holds $\mathbb{P}^{U_j}(p^{\text{rand.}}(\mathbf{x}_j, U_j) \leq \lambda) = 0$. Third, if $\lambda < p(\mathbf{x}_j) \leq \lambda + \sum_{\tilde{\mathbf{x}} \in A_j} f(\tilde{\mathbf{x}}|\mathbf{n})$, we easily calculate that

$$\mathbb{P}^{U_j}(p^{\text{rand.}}(\mathbf{x}_j, U_j) \leq \lambda) = 1 - \frac{p(\mathbf{x}_j) - \lambda}{\sum_{\tilde{\mathbf{x}} \in A_j} f(\tilde{\mathbf{x}}|\mathbf{n})}.$$

Altogether, this entails

$$\begin{aligned} \mathbb{E}^U[\hat{F}_M(\lambda)|\mathbf{X} = \mathbf{x}] &= \#\{1 \leq j \leq M : p(\mathbf{x}_j) \leq \lambda\}/M + \\ &\quad \sum_{j : \lambda < p(\mathbf{x}_j) \leq \lambda + \sum_{\tilde{\mathbf{x}} \in A_j} f(\tilde{\mathbf{x}}|\mathbf{n})} \left(1 - \frac{p(\mathbf{x}_j) - \lambda}{\sum_{\tilde{\mathbf{x}} \in A_j} f(\tilde{\mathbf{x}}|\mathbf{n})}\right) / M. \end{aligned}$$

Appendix IV: Remaining results for the replication study by Herder et al. (2008)

SNP	Allelic OR (adjusted)	one-sided p (adjusted)	Allelic OR (unadjusted)	one-sided $p^{\text{rand.}}$ (unadjusted)
rs10001190	0.89	0.09	0.88	0.0635
rs4458523	1.01	0.55	1.00	0.4805
rs4689394	0.99	0.45	0.98	0.4233
rs5018648	1.02	0.60	1.01	0.4251
rs10012946	1.00	0.52	0.99	0.4314
rs1046314	1.02	0.61	1.01	0.4268
rs564398	1.00	0.52	0.98	0.3882
rs7865618	0.97	0.36	0.96	0.2842
rs2383208	1.04	0.67	1.04	0.3735
rs10811661	1.09	0.80	1.08	0.2207
rs5015480	0.87	0.038	0.87	0.0352
rs10748582	0.84	0.022	0.86	0.0276
rs7923866	0.86	0.031	0.87	0.0369
rs7901695	1.21	0.010	1.23	0.0059
rs4506565	1.21	0.012	1.22	0.0078
rs4132670	1.22	0.0082	1.23	0.0055
rs7928810	1.07	0.20	1.08	0.1784
rs5215	1.08	0.16	1.09	0.1453
rs12790182	1.13	0.91	1.13	0.9165
rs1845618	1.10	0.85	1.10	0.8650
rs1113132	1.09	0.85	1.10	0.8672
rs7945827	1.09	0.83	1.09	0.8374
rs729287	1.10	0.86	1.11	0.8831
rs897004	1.04	0.67	1.04	0.6925
rs9939973	1.14	0.051	1.11	0.0933
rs9940128	1.14	0.053	1.11	0.0872
rs1121980	1.15	0.047	1.12	0.0861
rs7193144	1.11	0.095	1.09	0.1505
rs8050136	1.12	0.08	1.10	0.1177
rs9939609	1.10	0.11	1.08	0.1520
rs9930506	1.14	0.058	1.12	0.0683

Table 4: Remaining odds ratios and p-values for the replication study by Herder et al. (2008)

References

Agresti, A. (1992): "A survey of exact inference for contingency tables. With comments and a rejoinder by the author." *Stat. Sci.*, 7, 131–177.

Barras, L., O. Scaillet, and R. Wermers (2010): "False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas," *The Journal of Finance*, 65, 179–216.

Benjamini, Y. and D. Yekutieli (2001): "The control of the false discovery rate in multiple testing under dependency." *Ann. Stat.*, 29, 1165–1188.

Blanchard, G., T. Dickhaus, N. Hack, F. Konietschke, K. Rohmeyer, J. Rosenblatt, M. Scheer, and W. Werft (2010): "μTOSS - Multiple hypothesis testing in an open software system." *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 11, 12–19.

Cardillo, G. (2007): "Myfisher23: a very compact routine for fisher's exact test on 2x3 matrix," *http://www.mathworks.com/matlabcentral/fileexchange/15399*.

Cheverud, J. M. (2001): "A simple correction for multiple comparisons in interval mapping genome scans." *Heredity*, 87, 52–58.

Crans, G. G. and J. J. Shuster (2008): "How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial," *Stat Med*, 27, 3598–3611.

Evans, D. M., P. M. Visscher, and N. R. Wray (2009): "Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk," *Human Molecular Genetics*, 18, 3525–3531.

Finner, H., T. Dickhaus, and M. Roters (2007): "Dependency and false discovery rate: Asymptotics." *Ann. Stat.*,, 35, 1432–1455.

Finner, H. and V. Gontscharuk (2009): "Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses." *Journal of the Royal Statistical Society B*, 71, 1031–1048.

Finner, H. and K. Straßburger (2001a): "Increasing sample sizes do not always increase the power of UMPU-tests for 2×2 tables." *Metrika*, 54, 77–91.

Finner, H. and K. Straßburger (2001b): "UMP(U)-tests for a binomial parameter: A paradox." *Biometrical Journal*, 43, 667–675.

Finner, H. and K. Straßburger (2007): "A note on p-values for two-sided tests," *Biometrical Journal*, 49, 941–943.

Finner, H., K. Straßburger, I. M. Heid, C. Herder, W. Rathmann, G. Giani, T. Dickhaus, P. Lichtner, T. Meitinger, H.-E. Wichmann, T. Illig, and C. Gieger (2010): "How to link call rate and *p*-values for hardy-weinberg equilibrium as measures of genome-wide snp data quality," *Statistics in Medicine*,, 29, 2347–2358.

Fisher, R. A. (1922): "On the interpretation of χ^2 from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, 85, 87–94.

Gao, X., L. C. Becker, D. M. Becker, J. D. Starmer, and M. A. Province (2010): "Avoiding the High Bonferroni Penalty in Genome-Wide Association Studies." *Genetic Epidemiology*, 34, 100–105.

Gao, X., J. Starmer, and E. R. Martin (2008): "A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms." *Genetic Epidemiology*, 32, 361–369.

Habiger, J. D. and E. A. Peña (2011): "Randomised P-values and nonparametric procedures in multiple testing." *Journal of Nonparametric Statistics*, 23, 583–604.

Herder, C., W. Rathmann, K. Strassburger, H. Finner, H. Grallert, C. Huth, C. Meisinger, C. Gieger, S. Martin, G. Giani, W. A. Scherbaum, H. E. Wichmann, and T. Illig (2008): "Variants of the PPARG, IGF2BP2, CDKAL1, HHEX, and TCF7L2 genes confer risk of type 2 diabetes independently of BMI in the German KORA studies," *Horm. Metab. Res.*, 40, 722–726.

Howie, B. N., P. Donnelly, and J. Marchini (2009): "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet.*, 5, e1000529.

Langaas, M., B. H. Lindqvist, and E. Ferkingstad (2005): "Estimating the proportion of true null hypotheses, with application to DNA microarray data." *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 67, 555–572.

Lehmann, E. L. and J. P. Romano (2005): *Testing statistical hypotheses. 3rd ed.*, Springer Texts in Statistics. New York, NY: Springer.

Lewontin, R. C. and K. I. Kojima (1960): "The evolutionary dynamics of complex polymorphisms," *Evolution*, 14, 458–472.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis (2010): "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genet. Epidemiol.*, 34, 816–834.

Lydersen, S., M. W. Fagerland, and P. Laake (2009): "Recommended tests for association in 2 x 2 tables," *Stat Med*, 28, 1159–1175.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly (2007): "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nat. Genet.*, 39, 906–913.

McCarroll, S. A., F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. H. Shapero, P. I. de Bakker, J. B. Maller, A. Kirby, A. L. Elliott, M. Parkin, E. Hubbell, T. Webster, R. Mei, J. Veitch, P. J. Collins, R. Handsaker, S. Lincoln, M. Nizzari, J. Blume, K. W. Jones, R. Rava, M. J. Daly, S. B. Gabriel, and D. Altshuler (2008): "Integrated detection and population-genetic analysis of SNPs and copy number variation," *Nat. Genet.*, 40, 1166–1174.

Meinshausen, N., L. Meier, and P. Bühlmann (2009): "*p*-values for high-dimensional regression." *J. Am. Stat. Assoc.*, 104, 1671–1681.

Moskvina, V. and K. M. Schmidt (2008): "On multiple-testing correction in genome-wide association studies," *Genetic Epidemiology*, 32, 567–573.

Nyholt, D. R. (2004): "A simple correction for multiple testing for snps in linkage disequilibrium with each other." *Am. J. Hum. Genet.*, 74, 765–769.

Romano, J. P., A. M. Shaikh, and M. Wolf (2008): "Discussion: On methods controlling the false discovery rate," *Sankhyā*, 70, 169–176.

Rüschendorf, L. (2009): "On the distributional transform, Sklar's theorem, and the empirical copula process." *J. Stat. Plann. Inference*, 139, 3921–3927.

Sarkar, S. K. (2002): "Some results on false discovery rate in stepwise multiple testing procedures." *Ann. Stat.*, 30, 239–257.

Sarkar, S. K. (2008a): "On methods controlling the false discovery rate," *Sankhyā*, 70, 135–168.

Sarkar, S. K. (2008b): "Rejoinder: On methods controlling the false discovery rate," *Sankhyā*, 70, 183–185.

Schweder, T. and E. Spjøtvoll (1982): "Plots of *P*-values to evaluate many tests simultaneously." *Biometrika*, 69, 493–502.

Sen, P. K. (2008): "Discussion: On methods controlling the false discovery rate," *Sankhyā*, 70, 177–182.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004): "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach." *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 66, 187–205.

The 1000 Genomes Consortium (2010): "A map of human genome variation from population-scale sequencing," *Nature*, 467, 1061–1073.

The Wellcome Trust Case Control Consortium (2007): "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, 447, 661–678.

Wasserman, L. and K. Roeder (2009): "High-dimensional variable selection." *Ann. Stat.*, 37, 2178–2201.

Weir, B. S. (1996): *Genetic Data Analysis II.*, Sinauer Associates: Sunderland, MA.

Wigginton, J. E., D. J. Cutler, and G. R. Abecasis (2005): "A Note on Exact Tests of Hardy-Weinberg Equilibrium." *The American Journal of Human Genetics,*, 76, 887–893.

Willer, C. J., S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, S. C. Heath, N. J. Timpson, S. S. Najjar, H. M. Stringham, J. Strait, W. L. Duren, A. Maschio, F. Busonero, A. Mulas, G. Albai, A. J. Swift, M. A. Morken, N. Narisu, D. Bennett, S. Parish, H. Shen, P. Galan, P. Meneton, S. Hercberg, D. Zelenika, W. M. Chen, Y. Li, L. J. Scott, P. A. Scheet, J. Sundvall, R. M. Watanabe, R. Nagaraja, S. Ebrahim, D. A. Lawlor, Y. Ben-Shlomo, G. Davey-Smith, A. R. Shuldiner, R. Collins, R. N. Bergman, M. Uda, J. Tuomilehto, A. Cao, F. S. Collins, E. Lakatta, G. M. Lathrop, M. Boehnke, D. Schlessinger, K. L. Mohlke,

and G. R. Abecasis (2008): "Newly identified loci that influence lipid concentrations and risk of coronary artery disease," *Nat. Genet.*, 40, 161–169.

Ziegler, A. and I. R. König (2006): *A Statistical Approach to Genetic Epidemiology.*, Weinheim: Wiley.