Computational Statistics and Data Analysis 56 (2012) 1038–1051

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Modified versions of Bayesian Information Criterion for genome-wide association studies

Florian Frommlet [a,*], Felix Ruhaltinger [a], Piotr Twaróg [b], Małgorzata Bogdan [b]

[a] Department of Medical Statistics, Medical University Vienna, Austria

[b] Institute of Mathematics and Computer Science, Wrocław University of Technology, Poland

ARTICLE INFO

Article history:
Available online 17 May 2011

Keywords:
Genome-wide association
Multiple testing
Linear regression
Model selection
mBIC

ABSTRACT

For the vast majority of genome-wide association studies (GWAS) statistical analysis was performed by testing markers individually. Elementary statistical considerations clearly show that in the case of complex traits an approach based on multiple regression or generalized linear models is preferable to testing single markers. A model selection approach to GWAS can be based on modifications of the Bayesian Information Criterion (BIC), where some search strategies are necessary to deal with a huge number of potential models. Comprehensive simulations based on real SNP data confirm that model selection has larger power to detect causal SNPs in complex models than single-marker tests. Furthermore, testing single markers leads to substantial problems with proper ranking of causal SNPs and tends to detect a certain number of false positive SNPs, which are not linked to any of the causal mutations. This behavior of single-marker tests is typical in GWAS for complex traits and can be explained by an aggregated influence of many small random sample correlations between genotypes of the SNP under investigation and other causal SNPs. These findings might at least partially explain problems with low power and nonreplicability of results in GWAS. A real data analysis illustrates advantages of model selection in practice, where publicly available gene expression data as traits for individuals from the HapMap project are reanalyzed.

1. Introduction

Within the past five years genome-wide association studies (GWAS) have become an important tool for genetic scientists. There exist several excellent reviews which elucidate the statistical intricacies involved; see e.g. Balding (2006), Hirschhorn and Daly (2005), McCarthy et al. (2008) and Ziegler et al. (2008). The goal of GWAS is to detect genomic regions which are associated with some trait (either quantitative or categorical). Most GWAS are performed as case control studies with disease status as dichotomous trait. However, recently there has been growing interest in GWAS for quantitative traits, where both cross sectional as well as cohort studies have been performed.

Genome-wide association is based on the common disease–common variant assumption, which suggests that the occurrence of common complex diseases is influenced by a moderate number of (possibly interacting) disease alleles, the so-called causal variants. To find such causal variants genetic markers covering large parts of the genome are used. The most common type of markers are single nucleotide polymorphisms (SNPs). Current SNP array technology allows us to determine the state of up to one million SNPs within a single experiment.

The huge number of markers leads to a multiple testing problem which has been extensively discussed in the literature (see the discussion on this topic in Ziegler et al. (2008)). It is common practice in applied papers on GWAS to report

* Corresponding address: Spitalgasse 23, A-1090 Vienna, Austria. Tel.: +43 1 40400 7492; fax: +43 1 40400 7477.
E-mail address: Florian.Frommlet@meduniwien.ac.at (F. Frommlet).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.05.005

single-marker tests of SNPs. For a review on statistical tests for case control studies we refer again to Ziegler et al. (2008). Recommended significance levels to control family wise error are as small as $\alpha = 5 \cdot 10^{-8}$ (Dudbridge and Gusnanto, 2008), though occasionally larger significance levels like $\alpha = 10^{-6}$ are used. It is well understood that due to positive correlations between markers a simple Bonferroni correction is likely to be too conservative. Approaches to deal with correlation between SNPs include for example permutation tests like in Stranger et al. (2007).

In the fairly related area of QTL mapping based on designed breeding experiments the search over single markers was abandoned already quite a while ago in favor of multi-marker models. In this context the problem of selection of significant markers is equivalent to the choice of the "best" multiple regression or generalized linear model. This task is however rather difficult due to the large number of potential regressors. Specifically, in Broman and Speed (2002) it was noticed that classical model selection criteria like Akaike Information Criterion (AIC), and even Bayesian Information Criterion (BIC), tend to select too many markers. Addressing this problem Bogdan et al. (2004) introduced a modified version of BIC (mBIC), suited for the situation where a large number of markers is searched, but only relatively few markers are expected to be true signals. mBIC was motivated in a Bayesian setting, using informative priors on the model dimension, which prefer rather small models. In Bogdan et al. (2008a,b) it was observed that mBIC penalty is closely related to the multiple regression version of the Bonferroni correction for multiple testing. In a series of papers (Baierl et al., 2006, 2007; Bogdan et al., 2008a,b; Erhardt et al., 2010; Żak et al., 2007) based on simulation studies good properties of mBIC were documented. Recently some asymptotic optimality properties of various modifications of BIC have been shown (Frommlet et al., submitted for publication).

The primary aim of this article is to adopt the model selection approach based on modifications of BIC to GWAS. Specifically we make use of mBIC, a model selection criterion controlling the family wise error rate, and the recently proposed mBIC2 which controls the false discovery rate (FDR). In a comprehensive simulation study based on real SNP data performance of model selection is compared with single-marker tests. For the ease of presentation only ordinary regression models for quantitative traits are considered. To deal in particular with the computational burden of a simulation study on GWAS we devise a rather simple but efficient model search strategy. For single-marker tests the two most popular multiple testing procedures are applied, namely Bonferroni correction and the Benjamini–Hochberg procedure controlling FDR. It is demonstrated both theoretically and in the simulation study that model selection has larger power than single-marker tests to detect SNPs associated with complex traits. Perhaps even more important is the fact that single-marker tests have severe difficulties in ranking causal SNPs correctly. This problem can be explained by elementary statistical considerations, but as far as we know has not yet been discussed elsewhere.

The paper is organized as follows: Section 2.1 presents basic statistical facts about linear models and explains in detail why single-marker tests are suffering from rather low power in the case of complex models. Section 2.2 introduces modifications of BIC which are suitable for model selection in a high dimensional multiple regression setting. In particular mBIC and mBIC2 are presented, the primary model selection criteria used in this article. Competing model selection approaches for GWAS are discussed in Section 6.

In Section 3.1 a model search strategy is described which is particularly suited to the situation where among a huge number of markers only a small fraction is expected to be strongly associated with the trait. Section 3.2 describes the details of a simulation study based on real SNP data, whose results are presented in Section 4. Finally in Section 5 we reanalyze publicly available gene expression data (Stranger et al., 2007) as quantitative traits for the individuals genotyped in the HapMap project (The International HapMap Consortium, 2007).

2. Statistical preliminaries

2.1. Linear regression models

Let y_i, $i \in \{1, \ldots, n\}$, denote measurements of a quantitative trait in n individuals. Assume there are p SNPs where $p \gg n$, but only $k \ll n$ of them have an effect on y. Let $\mathbf{j}^\star = (j^\star_{l_1}, \ldots, j^\star_{l_k})$, with $1 \le j^\star_{l_1} < \cdots < j^\star_{l_k} \le p$, denote the ordered set of indexes of causal SNPs. We denote the genotype of person i and SNP j as $x_{ij} \in \{-1, 0, 1\}$ and assume that the following additive model holds:

$$\mathcal{M}_{\mathbf{j}^\star} : y_i = \beta_0 + \sum_{l=1}^{k} \beta_{j^*_l} x_{ij^*_l} + \epsilon_i. \tag{1}$$

For the sake of simplicity we assume that the error terms are i.i.d. normal random variables $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. In a more realistic scenario the model can be easily extended by including other covariates, like sex or age.

The model proposed in (1) is rather simple. However, complexity arises because the task is to find this model among 2^p possible models, where in GWAS p is of the order 10^7. The null model which does not include any causal SNP will be denoted by $\mathcal{M}_0$. All further additive models can be characterized by multi-indices $\mathcal{M}_{\mathbf{j}}$, where $\mathbf{j}$ is an ordered subset of elements of the set $\{1, \ldots, p\}$. Generically we will write $q = q_{\mathbf{j}}$ for the number of markers of a model. For the correct model we have $q_{\mathbf{j}^*} = k$.

Define for each model $\mathcal{M}_{\mathbf{j}}$ the matrix $X^{\mathbf{j}} = (\mathbf{1}, x_{j_1}, \ldots, x_{j_q})$, where $\mathbf{1} = (1, \ldots, 1)'$. Then we have in vector notation

$$\mathcal{M}_{\mathbf{j}} : y = X^{\mathbf{j}} \beta_{\mathbf{j}} + \epsilon^{\mathbf{j}}, \tag{2}$$

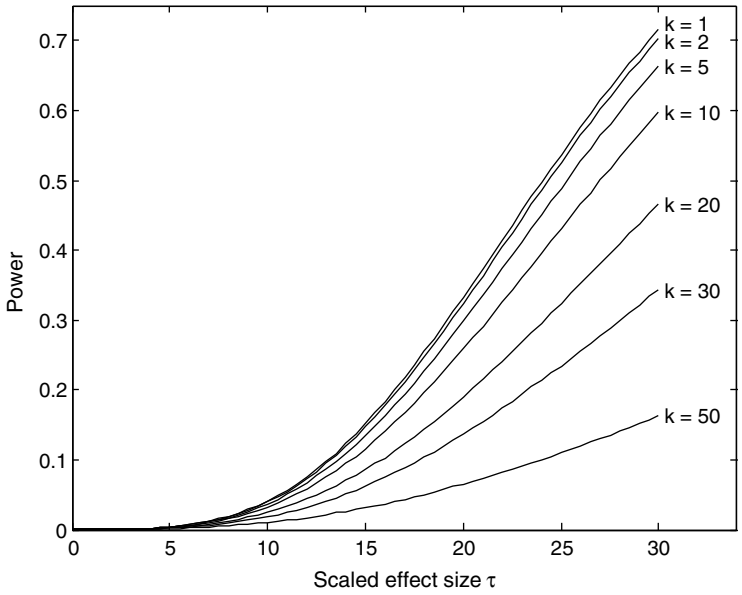

Fig. 1. Power to find a causal SNP with single-marker tests when model $\mathcal{M}_{\mathbf{j}^*}$ with k effects is correct. In the case of $k = 1$ one is testing for the correct model.

where $\beta_{\mathbf{j}} = (\beta_0, \beta_{j_1}, \ldots, \beta_{j_q})'$. Given the large number of markers it is understandable that it is common practice to perform only single-marker analysis. This means that only models are considered which are of the form

$$\mathcal{SM}_j : y_i = \beta_0 + \beta_j x_{ij} + \epsilon_i^{(j)}. \tag{3}$$

Elementary statistics tells us what happens when we perform an F-test for some model $\mathcal{M}_{\mathbf{j}}$ based on least squares regression. Let $P_{\mathbf{j}} = (X^{\mathbf{j}})'[X^{\mathbf{j}}(X^{\mathbf{j}})']^{-1}X^{\mathbf{j}}$ denote the usual hat-operator for a general model $\mathcal{M}_{\mathbf{j}}$. Then $RSS_{\mathbf{j}} := y'(I - P_{\mathbf{j}})y$ is the residual sum of squares and $MSS_{\mathbf{j}} := y'(P_{\mathbf{j}} - \frac{1}{n}E)y$ is the model sum of squares for $\mathcal{M}_{\mathbf{j}}$. We always denote by I the identity matrix and by $E = \mathbf{1}\mathbf{1}'$ the all one matrix of suitable dimension (here they are $n \times n$). The usual F-test statistic for the null hypothesis that none of the variables in the model $\mathcal{M}_{\mathbf{j}}$ has an influence on Y is given by

$$F_{\mathbf{j}} = \frac{(n - q_{\mathbf{j}} - 1)MSS_{\mathbf{j}}}{q_{\mathbf{j}}RSS_{\mathbf{j}}}.$$

When the model $\mathcal{M}_{\mathbf{j}}$ includes all causal SNPs then the statistics $F_{\mathbf{j}}$ has a noncentral F-distribution and power calculations for different effect sizes are rather straight forward (compare results for $k = 1$ in Fig. 1). However, we are often facing a different situation when model $\mathcal{M}_{\mathbf{j}^*}$ holds, but we are performing an F-test for a smaller model $\mathcal{M}_{\mathbf{j}}$, which might not include some of the causal SNPs. Then

$$\epsilon_i^{\mathbf{j}} = y_i - \beta_0 - \sum_{l=1}^{q_j} \beta_{j_l} x_{ij_l} \sim \mathcal{N}\left(\sum_{l \in \mathbf{j}^* \setminus \{\mathbf{j}\}} \beta_l x_{il}, \sigma^2\right),$$

and according to the generalization of Cochran's theorem for the noncentral case (Madow, 1940) the model sum of squares and the residual sum of squares are independent with distributions

$$\frac{MSS_{\mathbf{j}}}{\sigma^2} \sim \chi^2\left(q_{\mathbf{j}}, \frac{1}{\sigma^2}\beta'_{\mathbf{j}^*}(X^{\mathbf{j}^*})'\left(P_{\mathbf{j}} - \frac{1}{n}E\right)X^{\mathbf{j}^*}\beta^*_{\mathbf{j}}\right), \tag{4}$$

$$\frac{RSS_{\mathbf{j}}}{\sigma^2} \sim \chi^2\left(n - q_{\mathbf{j}} - 1, \frac{1}{\sigma^2}\beta'_{\mathbf{j}^*}(X^{\mathbf{j}^*})'(I - P_{\mathbf{j}})X^{\mathbf{j}^*}\beta_{\mathbf{j}^*}\right). \tag{5}$$

Here $\chi^2(u, v)$ denotes a noncentral chi-square distribution, with the number of degrees of freedom equal to u and a noncentrality parameter v. Thus the test statistic $F_{\mathbf{j}}$ is essentially the ratio of two independent noncentral χ^2-distributed random variables. If the size of the true model $q_{\mathbf{j}^*}$ is much larger than the size $q_{\mathbf{j}}$ of the model under consideration, then the residual sum of squares $RSS_{\mathbf{j}}$ will have a considerably large noncentrality parameter incorporating effects which have not entered the model, and the power of the according F-test will be comparably small.

This effect will be most pronounced in the case of simple regression models (3). To fix ideas consider for a moment the orthogonality assumption $(X^{\mathbf{j}^*})'X^{\mathbf{j}^*} = nI$. Then the noncentrality parameters corresponding to MSS_j and RSS_j respectively

become

$$\nu_{M,j} = \frac{n\beta_j^2}{\sigma^2}, \quad \text{and} \quad \nu_{R,j} = \sum_{l \in \mathbf{j}^* \setminus \{j\}} \frac{n\beta_l^2}{\sigma^2}. \tag{6}$$

In Fig. 1 power calculations are shown for this simplified situation with $n = 2000$ and $\alpha = 10^{-6}$. The squared scaled effect sizes $\tau = \frac{n\beta_l^2}{\sigma^2}$ are equal for all k effects. Power was obtained by sampling from the two noncentral χ^2-distributions of (4) and (5). If there is only a small number of causal SNPs the loss of power by testing for individual markers is not dramatic. However already for $k = 10$ the loss becomes recognizable, and for $k = 30$ one is actually losing more than 50% of power in the range of effect sizes we considered.

Now in GWAS one can certainly not expect that all causal SNPs have the same effect size, and their genotypes will also not be orthogonal. However, GWAS are performed to understand the genetics of complex traits, which per definition are influenced by more than one factor. Therefore our considerations concerning loss of power by single-marker analysis will apply. For a single effects model $\mathcal{SM}_j$ one obtains

$$(I - P_j)X^{\mathbf{j}^*}\beta_{\mathbf{j}^*} = \sum_{l \in \mathbf{j}^* \setminus \{j\}} \beta_l \left((x_l - \bar{x}_l) - \frac{\mathrm{Cov}(x_j, x_l)}{\mathrm{Var}(x_j)} (x_j - \bar{x}_j) \right),$$

where $\bar{x}_j$ is the sample mean and $\mathrm{Var}(x_j)$ and $\mathrm{Cov}(x_j, x_l)$ are the sample variance and covariance respectively. The noncentrality parameters for single-marker tests have the form

$$\nu_{M,j} = \frac{\left(\sum_{l=1}^{k} \beta_l \mathrm{Cov}(x_j, x_l) \right)^2}{\sigma^2 \mathrm{Var}(x_j)}, \tag{7}$$

and

$$\nu_{R,j} = \sum_{l \in \mathbf{j}^* \setminus \{j\}} \sum_{r \in \mathbf{j}^* \setminus \{j\}} \frac{\beta_l \beta_r}{\sigma^2} \left(\mathrm{Cov}(x_l, x_r) - \frac{\mathrm{Cov}(x_l, x_j)\mathrm{Cov}(x_r, x_j)}{\mathrm{Var}(x_j)} \right). \tag{8}$$

Compared to the case of orthogonality in (6) things are slightly more complicated due to correlation effects, which might have a strong influence on the noncentrality parameters for $RSS_{\mathbf{j}}$ and $MSS_{\mathbf{j}}$. This issue will be addressed in detail in Section 3.2.

2.2. Modifications of BIC

Assume that a family of models $\mathcal{M}_{\mathbf{j}}$ has parameters $\theta_{\mathbf{j}}$ and corresponding likelihood functions $L_{\mathbf{j}}(\theta_{\mathbf{j}})$. Denote by $\hat{\theta}_{\mathbf{j}}$ the maximum likelihood estimates of $\theta_{\mathbf{j}}$. Many statistical model selection criteria, like for example AIC or BIC, suggest to select that model which maximizes a penalized likelihood function of the form

$$\log L_{\mathbf{j}}(\hat{\theta}_{\mathbf{j}}) - \eta\, q_{\mathbf{j}}. \tag{9}$$

For AIC and BIC the penalty parameter η takes the form 1 and $\frac{1}{2}\log n$ respectively.

For linear regression under the assumption of normal error terms $\epsilon^{\mathbf{j}} \sim \mathcal{N}(0, \sigma^2 I)$ the likelihood function of each model $\mathcal{M}_{\mathbf{j}}$ in (2) is given by

$$L_{\mathbf{j}}(y|\beta_{\mathbf{j}}, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{(y - X^{\mathbf{j}}\beta_{\mathbf{j}})'(y - X^{\mathbf{j}}\beta_{\mathbf{j}})}{2\sigma^2} \right).$$

The maximum likelihood estimator of $\beta_{\mathbf{j}}$ then coincides with the least squares regression estimator $\hat{\beta}_{\mathbf{j}}$ and thus

$$L_{\mathbf{j}}(y|\hat{\beta}_{\mathbf{j}}, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{RSS_{\mathbf{j}}}{2\sigma^2} \right).$$

For fixed σ BIC is then equivalent to minimize

$$\frac{RSS_{\mathbf{j}}}{\sigma^2} + q_{\mathbf{j}} \log n. \tag{10}$$

For unknown σ the ML-estimate $\hat{\sigma}^2 = \frac{RSS_{\mathbf{j}}}{n}$ leads to the criterion

$$n \log RSS_{\mathbf{j}} + q_{\mathbf{j}} \log n. \tag{11}$$

For sample size $n \geq 8$ the penalty parameter in AIC is smaller than in BIC and therefore in this case BIC tends to select more parsimonious models than AIC. Also, it is known that when p is fixed and n goes to infinity then BIC is consistent. Thus,

when n is large and $p \ll n$, BIC usually selects the true model with large probability. However, the situation is very different in the case of $p > n$. As explained in detail in Bogdan et al. (2008b) BIC is derived with the underlying prior assumption that all possible models $\mathcal{M}_{\mathbf{j}}$ are chosen with the same probability. This effectively results in using a Binomial prior $B(p, 1/2)$ on the model dimension. Thus, BIC assigns a high prior probability to the class of models of size $p/2$, whereas small or very large dimensions are much less likely a priori. Under sparsity, where the actual model has only a small number of regressors, this results in BIC choosing too many regressors.

As a remedy for this situation Bogdan et al. (2004) introduced a modification of BIC, which can be formulated as

$$\text{mBIC: } -2\log L_{\mathbf{j}}(\hat{\theta}_{\mathbf{j}}) + q_{\mathbf{j}}(\log n + 2\log p + d). \tag{12}$$

This criterion was derived in a Bayesian setting assuming a prior probability of the model $\mathcal{M}_{\mathbf{j}}$ of the form

$$\pi(\mathbf{j}) = \omega^{q_{\mathbf{j}}}(1-\omega)^{p-q_{\mathbf{j}}}.$$

In our context ω can be interpreted as the a priori expected proportion of causal SNPs, where all SNPs have independently from each other the same chance of being causal. This is a typical prior assumption in Bayesian model selection (see e.g. Chipman et al. (2001)). Incorporating this prior distribution into BIC we easily obtain (12) with $d = -2\log(p\omega)$, i.e. minus two times the logarithm of the expected number of causal SNPs (for details of this derivation see Bogdan et al. (2004)). Without any prior knowledge on the expected number of SNPs a standard choice for d is $-2\log 4$, as motivated in Bogdan et al. (2008b).

In the case of known σ and under the assumption of orthogonal regressors mBIC has been shown to be closely related to the Bonferroni correction rule for multiple testing (Bogdan et al., 2008b). In particular mBIC is controlling the family wise error. In Frommlet et al. (submitted for publication) it is shown that under certain sparsity conditions mBIC is consistent and has some optimality properties. Furthermore, mBIC has been studied in the context of Generalized Linear Models (Żak-Szatkowska and Bogdan, in press) as well as Zero Inflated Generalized Poisson Regression (Erhardt et al., 2010).

Recently, Chen and Chen (2008) proposed a new modification of BIC called extended Bayesian Information Criterion (EBIC). EBIC assigns a prior probability for the model dimension q, which is proportional to $\binom{p}{q}^{\kappa}$, for some $\kappa \in [0, 1]$. This results in the criterion

$$\text{EBIC: } -2\log L_{\mathbf{j}}(\hat{\theta}_{\mathbf{j}}) + q_{\mathbf{j}}\log n + 2\log\binom{p}{q}^{1-\kappa}.$$

If $\kappa = 1$ then EBIC coincides with BIC. The choice $\kappa = 0$ corresponds to the uniform prior on the model dimension. In Chen and Chen (2008) some consistency properties of EBIC are proved under the assumptions that the maximal dimension searched by EBIC is fixed and larger than the true number of effects. In Chen and Chen (submitted for publication) EBIC was further extended to Generalized Linear Models and in Zhao and Chen (submitted for publication) it was successfully used for GWAS with binary traits.

While EBIC turns out to work very well in many practical sparse cases, it has one undesirable property. When $q > \frac{p}{2}$ the last term of the penalty becomes a decreasing function of q and encourages to pick the largest possible model. Therefore in this article we will consider a slightly different criterion,

$$\text{mBIC2: } -2\log L_{\mathbf{j}}(\hat{\theta}_{\mathbf{j}}) + q_{\mathbf{j}}(\log n + 2\log p + d) - 2\log(q_{\mathbf{j}}!), \tag{13}$$

which is asymptotically equivalent to EBIC with $\kappa = 0$ when the maximal allowable number of regressors, Q, is of the order $Q = o(p)$. Again the choice $d = -2\log 4$ is recommended, see Frommlet et al. (submitted for publication) for details.

mBIC2 was developed by Frommlet et al. (submitted for publication) as a model selection rule which in the context of multiple regression controls FDR, working similarly to the Benjamini and Hochberg (1995) correction for multiple testing. In Frommlet et al. (submitted for publication) a thorough discussion is provided how this modification of mBIC relates to a similar criterion suggested by Abramovich et al. (2006) as well as to a modification of the risk inflation criterion RIC, proposed in George and Foster (2000). Due to the negative extra term mBIC2 will potentially select larger models than the original mBIC (12). In Frommlet et al. (submitted for publication) it is shown that mBIC2 has asymptotic optimality properties for a much larger range of sparsity levels than the original mBIC. We will compare the behavior of both criteria to select causal SNPs in the simulation study of Section 3.2.

3. Methods

3.1. Search algorithm

An important question when applying a model selection approach to GWAS data is how to deal with the large number of possible models. Some interesting search strategies for the best multiple regression model were recently proposed e.g in Zhao and Chen (submitted for publication) and Chen and Chen (2009). However, these advanced model selection strategies are rather unfeasible for large scale simulation studies. Therefore, for the purpose of our simulation study we developed our own search strategy, whose initial step relies on some modification of the popular forward selection. Our method takes into

account the fact that we are expecting a rather moderate number of causal SNPs (somewhere below 100) and turns out to be relatively accurate and fast enough to allow for a simulation study based on more than 300000 SNPs.

In an initial step we perform single-marker tests for all SNPs, a step we have to take in any case to be able to compare our model selection approach with the Bonferroni and Benjamini–Hochberg (BH) procedures. For further analysis we only consider SNPs with an uncorrected p-value smaller than 0.15.

The second step consists of a simplified forward search strategy which we call multiple forward search. To this end we start with computing the original BIC (11) for the single-marker model with lowest p-value. Then we proceed iteratively by considering SNPs in ascending order of single-marker p-values and decide based on BIC to enhance the current model by a new SNP or not. This procedure is performed till we have considered all SNPs, or we have reached a maximum model size of 140 SNPs. For practical reasons we do not allow for larger models at this stage.

The initial multiple forward selection, based on the uncorrected BIC, is expected to include a lot of false positives in the model, but hopefully also many causal SNPs. Its principal advantage is that the actual search procedure, based on modifications of BIC, is not starting from the null hypothesis, but from a large model for which the residual sum of squares will have been considerably reduced. From here we start to perform backward selection and then stepwise selection based either on mBIC (12) or on mBIC2 (13). This search strategy is designed to overcome the difficulties discussed in Section 2.1, when the actual model includes a large number of causal SNPs. It works well in the simulation study of Section 3.2, where more time consuming search procedures are out of question. In the real data analysis of Section 5 the procedure will be amended with a final step, where all subsets of a specified set of SNPs are considered.

3.2. Simulation study

Simulations are based on SNP data from the population reference sample POPRES (Nelson et al., 2008). The data set used for simulations in this manuscript was obtained from dbGaP through dbGaP accession number phg000027.v2.p2 at www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000145.v1.p1. In particular we used data from the file Glaxo.txt, which contains a subsample of individuals studied in the article Lao et al. (2008). It comprises genotypes from 309788 SNPs for 649 individuals, which all have European ancestry and represent a relatively homogeneous population.

In this data set approximately 5% of genotype values were missing. To deal with these missing values we adopted the following imputation strategy. Suppose x_{ij}, the genotype of SNP j for the ith individual, is missing. We search for the 4 SNPs with strongest correlation to SNP j fulfilling two conditions: They are in a neighborhood of 500 SNPs upstream or downstream of SNP j and their values for the ith individual are not missing. If we find individuals who have exactly the same values as the ith individual on these 4 SNPs, then we predict the value of x_{ij} as the most frequent value of SNP j among these individuals. If we cannot find individuals fulfilling the above mentioned condition, then the most frequent value of the jth SNP among all individuals is imputed. Since the main purpose of this experiment is to present some basic properties of different approaches to GWAS, both our simulation and search procedures treat the final set of obtained SNP genotypes as the "correct" one. Therefore, the imputation procedure has no influence on the final results.

Among the $p = 309788$ SNPs we have chosen $k = 40$ SNPs from autosomal chromosomes to be causal, meaning they comprise the regressors of the linear model under which values of the quantitative trait are simulated. These were selected deliberately in such a way that they are common and well distributed over all chromosomes. The minimum allelic frequency for all causal SNPs was ranging between 0.3 and 0.5; variance of their genotype data was ranging between 0.42 and 0.53; and correlations between all possible pairs of causal SNPs were between -0.12 and 0.1. For the considered sample size this range of sample correlations corresponds well to the range of random sample correlations between independent SNPs.

We simulated 1000 replicates from the additive model (1) where $\mathbf{j}^*$ indicates the 40 causal SNPs. Model selection procedures are much more time consuming than standard procedures based on multiple testing, and it seems that 1000 replicates is actually rather large compared with numbers found in the literature. For example Hoggart et al. (2008) used 500 replicates, He and Lin (2011) used only 200 replicates, and Li et al. (2010) even only 100 replicates. Note that the standard error of the estimated power based on 1000 replicates is rather small (1.5% when the power is equal to 50% and even smaller in other cases).

Error terms were sampled from a standard normal distribution, i.e. $\sigma^2 = 1$. The 40 effect sizes were equally distributed between 0.27 and 0.66. The overall heritability, defined as

$$H^2 = \frac{\mathrm{Var}(X^{\mathbf{j}^*}\beta_{\mathbf{j}^*})}{1 + \mathrm{Var}(X^{\mathbf{j}^*}\beta_{\mathbf{j}^*})}, \tag{14}$$

is equal to 0.81. Heritability of an individual effect, defined as

$$h^2_{j^*} = \frac{\beta^2_{j^*}\,\mathrm{Var}(x_{j^*})}{1 + \mathrm{Var}(X^{\mathbf{j}^*}\beta_{\mathbf{j}^*})}, \tag{15}$$

ranges between 0.006 and 0.037. We are aware of the fact that the overall heritability is unrealistically large, but we consider it instructive to present the difficulties of the multiple testing procedures, which occur even in this simplified setting.

Table 1
Four different selection procedures used in the simulation study.

	Single-marker analysis	Model selection
Controlling FWER	Bonferroni	mBIC
Controlling FDR	Benjamini–Hochberg	mBIC2

Each simulated data set was analyzed using the two most popular multiple testing procedures (Bonferroni and Benjamini–Hochberg) as well as model selection approaches based on mBIC and mBIC2. In this way it is possible to compare differences between single-marker analysis and model selection, while at the same time comparing procedures controlling the family wise error rate with FDR controlling procedures (see Table 1). Bonferroni multiple testing correction was performed at family wise error rate $\alpha = 0.05$, which corresponds to an adjusted significance level of approximately $1.6 \cdot 10^{-7}$. Benjamini–Hochberg procedure was performed at the corresponding FDR level $\alpha = 0.05$. Model selection with mBIC was based on the constant $d = -2\log(4)$, which serves as a standard choice (see e.g. Bogdan et al. (2008b)). Based on the calculations of Bogdan et al. (2004) we expect that for p and n of this data set mBIC controls the family wise error under the total null approximately at a level $\alpha = 0.02$. We have also computed Bonferroni correction and BH at this smaller level, but given the observed lack of power of both BH and Bonferroni these results are not presented.

In GWAS studies it frequently happens that not the causal SNP itself is detected as significant, but some SNP whose genotype is highly correlated to the causal SNP. Such a finding is not necessarily to be considered as a false positive, which leads to the question how to define true and false positives for correlated regressors. We adopt the following convention: Let genotypes be coded as ± 1 for homozygous SNPs and 0 for heterozygous SNPs. The correlation between two SNPs is defined as the correlation between the corresponding SNP data vectors. Any detected SNP whose correlation to a causal SNP has absolute value larger than a given threshold is counted as a true positive (TP), otherwise as a false positive (FP). We initially used a threshold of $|R| = 0.9$, and based on simulation results decided to report alternatively also results for thresholds $|R| = 0.7$ and $|R| = 0.5$. Here $|R|$ is the maximum absolute value over all correlations with causal SNPs.

For multiple testing procedures it often occurs that two or more selected SNPs are correlated with a causal SNP. In case they are above the specified threshold they are all counted as just one true positive. False positive SNPs with identical genotype are only counted once. Based on these conventions for each simulated data set the false discovery rate (FDR) is computed as the percentage of false positives among all detections (0 if there are no detections) and the average power is defined as the percentage of detected causal SNPs. Furthermore, the power for each individual causal SNP is estimated as the detection rate over all 1000 replicates.

4. Results

The boxplots in Fig. 2 illustrate the observed FDR values and average power for the 1000 simulated replicates. It is evident that mBIC2 has the largest average power among the 4 different procedures. Also, as expected, mBIC has larger average power than the two multiple testing procedures. On the other hand both multiple testing procedure have larger FDR than mBIC2, which itself has larger FDR than mBIC. Both model selection procedures show much less variation in FDR than BH, which is a direct consequence of the fact that the model selection procedures have larger power.

Comparing procedures controlling FDR and FWER with each other, it is no surprise that mBIC2 has considerably larger power than mBIC, while BH has larger power than Bonferroni correction. The difference in FDR strongly depends on the choice of the threshold value of $|R|$ which determines a "false positive".

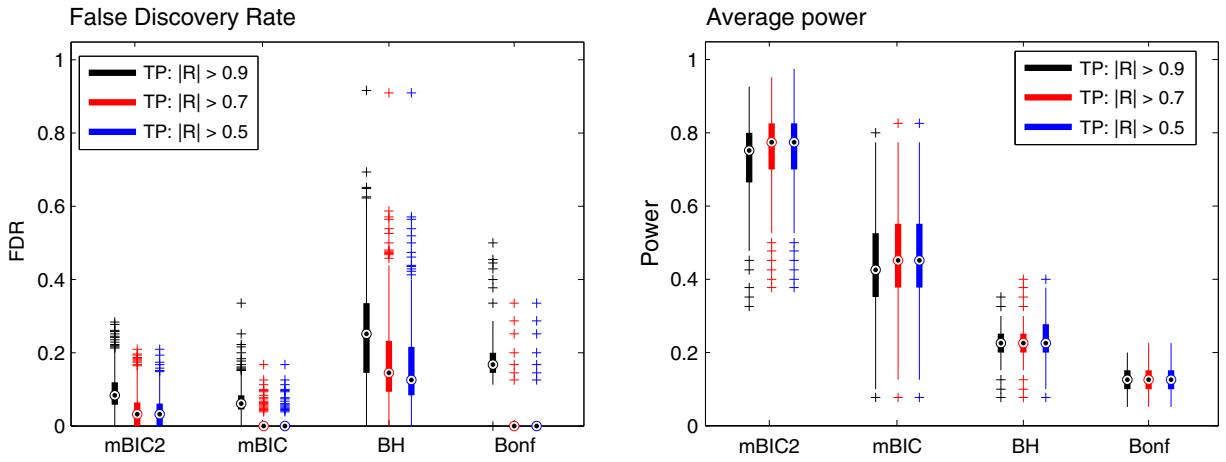

Fig. 2. FDR and average power for the 4 different selection procedures. For each procedure three boxplots are shown. A detected SNP was classified as a true positive when its maximal correlation to a causal SNP was larger than 0.9 on the left, 0.7 in the middle and 0.5 on the right.

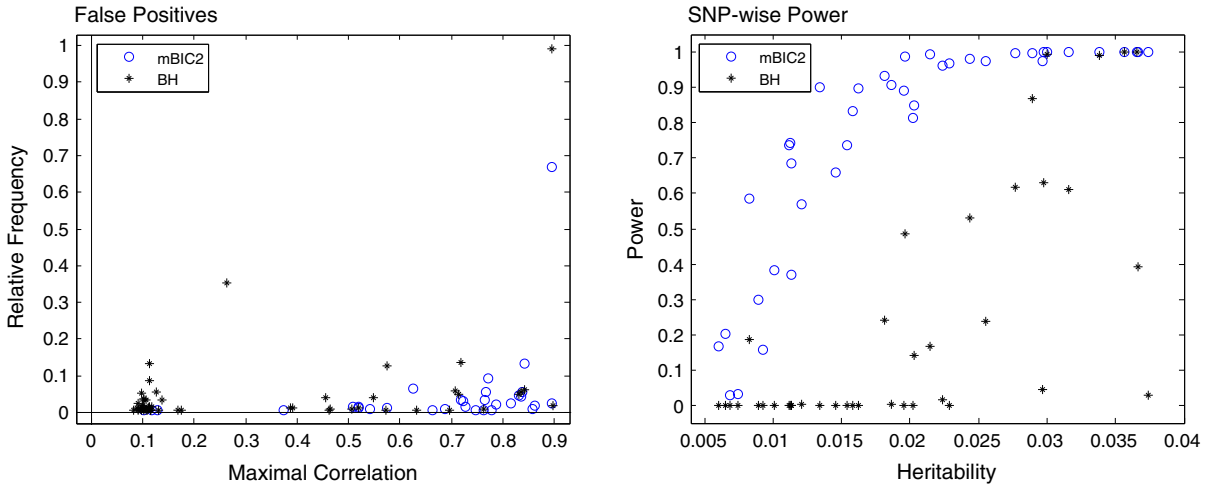

Fig. 3. First plot: Relative frequency of false positive SNPs against their maximal correlation with any of the 40 causal SNPs. Only SNPs are shown which have been detected in at least five simulation runs as false positives. Second plot: Power to detect each causal SNP as a function of the individual heritability.

4.1. Dependence of TP and FP on |R|-thresholds

In Fig. 2 FDR and average power were computed for each method based on thresholds $|R| = 0.9$, $|R| = 0.7$ and $|R| = 0.5$, respectively. At a threshold level $|R| = 0.9$ differences in FDR between mBIC and mBIC2 or between BH and Bonferroni are not that large. For all procedures FDR is noticeably reduced when using the more liberal criterion $|R| = 0.7$. In particular mBIC and the Bonferroni procedure have in that case considerably less false positives. The change from $|R| = 0.7$ to $|R| = 0.5$ is less dramatic.

We will from now on focus on results for mBIC2 and BH. The first plot in Fig. 3 shows how often certain SNPs occur as false positives based on the threshold $|R| = 0.9$. Only those SNPs are plotted which were detected in at least five simulation runs as false positives. The corresponding tables presenting these results in detail are provided in Web_Supplement_3.txt. The first significant finding is that SNP A-2299101, detected in 670 simulation runs by mBIC2, has a correlation of 0.8958 with causal SNP A-1912140. This explains the low power of 33% to detect causal SNP A-1912140 with mBIC2, and we conclude that SNP A-2299101 is actually not really a false positive, but rather it is detected instead of SNP A-1912140. Thus the threshold $|R| = 0.9$ to determine false positives is apparently too strict.

Looking at the first Table of Web_Supplement_3.txt we observe that all SNPs detected by mBIC2 in more than 5 simulation runs have $|R| > 0.5$. Only four SNPs which were detected 5 times had $|R| < 0.5$. Since none of the statistical approaches to GWAS can clearly distinguish SNPs which are closely correlated, we believe that instead of reporting just one detected SNP, one should report also all SNPs which are strongly correlated to it. The smaller the applied threshold the more SNPs one has to report, and to this end the value 0.5 appears to be fairly small. As a compromise we decided to consider $|R| = 0.7$ as a suitable threshold, though even smaller thresholds have been suggested in the literature (Hoggart et al., 2008). With the exception of SNP A-4270622 all SNPs which are detected by mBIC2 in at least 18 simulation runs are then classified as true positives.

On the other hand there is a relatively large number of SNPs which are detected by mBIC2 only once (1076), twice (55) or three times (12). The majority of these detections, which might be classified as actual false positives, are usually not correlated to any causal SNP ($|R| < 0.5$ held for 1050 SNPs detected once, for 48 detected twice and for 9 detected three times). Based on a model selection approach one would expect to detect some false positives of this nature, and their frequency is controlled according to the theory of mBIC2.

Looking at the first plot of Fig. 3, it is a certain peculiarity that the frequency of several false positive SNPs detected by BH is peaking for rather small values of $|R|$. One example is SNP A-2181789, which has been detected 354 times as a false positive by BH, but is only correlated with $|R| = 0.2628$ to the closest causal SNP. Several other frequently detected SNPs have maximal correlation way below 0.2, for details see the second Table of Web_Supplement_3.txt. An explanation for this phenomenon is given in the next section.

4.2. Single-marker analysis and heritability

The second graph in Fig. 3 provides the estimated power to detect each of the 40 causal SNPs at a threshold level $|R| = 0.7$ as a function of the individual heritability (15). For mBIC2 one observes the expected trend that larger individual heritability yields larger power, whereas for BH the behavior is much more erratic. The order of SNPs with respect to power is quite different from the order in terms of individual heritability. For example, the SNP with the largest heritability, easily detected

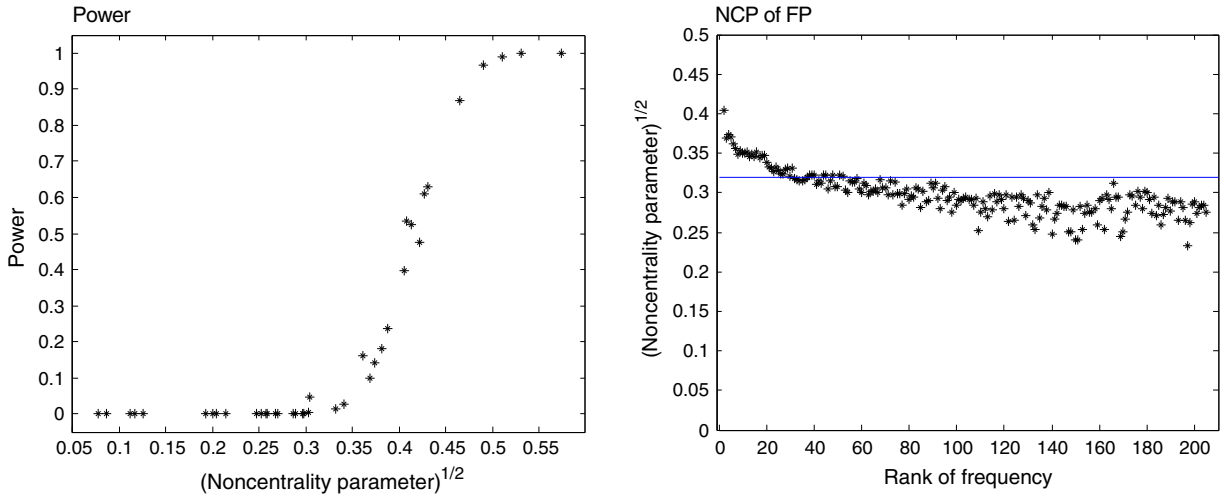

Fig. 4. First plot: Power of the Benjamini–Hochberg procedure to detect causal SNPs plotted against the square root of the noncentrality parameter instead of the individual heritability. Second plot: Square root of the noncentrality parameter for all false positives occurring under BH (at level $|R| = 0.9$). On the x-axis SNPs are ordered in decreasing order according to their frequency of detection in the 1000 simulation runs. Stars with lowest rank correspond to the SNPs listed in the second Table of Web_Supplement_3.txt.

by mBIC2, is completely missed by the Benjamini–Hochberg procedure. On the other hand, some substantially "weaker" SNPs, are detected with a power exceeding 50%.

Remember that F-tests of single effect models involve noncentrality parameters (7) and (8). We can rewrite the square root of (7) as

$$\sqrt{\nu_{M,j}} = \left| \frac{\beta_j}{\sigma} \sqrt{\mathrm{Var}(x_j)} + \frac{\sum_{l \neq j} \beta_l \,\mathrm{Cov}(x_j, x_l)}{\sigma \sqrt{\mathrm{Var}(x_j)}} \right| .$$

This shows that in the case of an orthogonal design matrix $\nu_{M,j}$ is proportional to the individual heritability. However, in the general case the noncentrality parameter is modified according to $\sum_{l \neq j} \beta_l \,\mathrm{Cov}(x_j, x_l)$. This term can occasionally become fairly large when there is a large number of true signals. Note that we designed our simulation study such that pairwise correlation between SNPs was small. In a statistical sense the genotypes of the causal SNPs can be thought of as being independent. Still, for some causal SNPs the effects of correlation just by chance accumulate significantly.

The first plot in Fig. 4 shows that these small pairwise correlations with other causal SNPs explain the erratic behavior of the Benjamini–Hochberg procedure. Plotting the power against the square root of the noncentrality parameter $\nu_{M,j}$ one observes the regular behavior of a sigmoid function. Clearly not the individual heritability but the weighted sum of correlations to all causal SNPs from (7) determines the power to detect an individual SNP. This observation is crucial. It calls into question the practice of reporting detected SNPs according to the order of p-values from multiple testing procedures and claiming that SNPs with smallest p-values are the most important ones. It might as well be the case that such signals are just catching the effect of many other causal SNPs which themselves are not detectable. Also, since most of the detected pairwise correlations between "false" SNPs and the trait result only from random fluctuations of sample correlation coefficients between these SNPs and the causal ones, they are not replicable in different samples from the same population. Therefore, such "false positives" are useless also in terms of predicting the trait values.

The second plot in Fig. 4 gives the answer to the question why some SNPs occur so frequently as false positives when they are not at all correlated with any causal SNP. It turns out that all false positive SNPs under BH have relatively large noncentrality parameter ($\sqrt{\nu_{M,j}} > 0.23$), and in particular those SNPs having been detected at least 5 times, i.e. those listed in the second table of Web_Supplement_3.txt, all have $\sqrt{\nu_{M,j}} > 0.32$. This is just the onset of the sigmoid function observed in the first plot of Fig. 4 where the power to detect causal SNPs no longer vanishes.

The conclusion from this analysis is the following. If we believe that a trait in a genome-wide association study is influenced by a relatively large number of genes, then single-marker analysis has a large chance of missing many of these genes. On the other hand there is a large chance of detecting false positive SNPs which have nothing to do with functional regions. This appears to be quite a devastating résumé for the performance of single-marker analysis in GWAS with complex traits.

5. Real data analysis

Stranger et al. (2007) have analyzed the association between SNPs from the HapMap project (The International HapMap Consortium, 2007) and gene expression data. They considered 270 individuals from four populations, namely 30 Caucasian

trios of northern and western European background (CEU), 30 Yoruba trios from Ibadan, Nigeria (YRI), 45 unrelated individuals from Beijing, China (CHB) and 45 unrelated individuals from Tokyo, Japan (JPT). The major objective of Stranger et al. (2007) was to find so called *cis* associations and *trans* associations between SNPs and gene expression, where the cis region for SNPs was defined to be 1 Mb upstream or downstream of the expression probe midpoint.

For statistical analysis (Stranger et al., 2007) used a permutation test approach for test statistics obtained with simple linear regression models only considering additive effects. Excluding the 60 children from the CEU and YRI trios left 210 unrelated individuals. Analysis was performed considering the four populations separately, as well as combining data from individuals of certain populations. Pooling over all four populations provides the most powerful approach, and we will thus restrict our analysis to this situation. To deal with population structure (Stranger et al., 2007) used a procedure based on conditional permutations, which was originally described in Koren et al. (2006).

We want to compare our model selection approach in particular with the analysis of Stranger et al. (2007) for trans association of gene expression with SNPs. Pooling all four populations they found 44 genes showing trans association, where detailed results are provided in their Supplementary Table 6. To make our results comparable with Stranger et al. (2007) we also restrict ourselves to additive models of the form (1), though we believe it would be interesting to consider additionally dominance effects. To account for population structure in our models we add dummy variables for populations CEU, CHB and JPT.

In Stranger et al. (2007) only some 25000 candidate SNPs were considered as putatively functional SNPs. In contrast we extend our search over all available SNPs. As a consequence mBIC2 will use a much larger penalty for multiple testing making it more difficult to include SNPs found by Stranger et al. (2007) in our model. On the other hand functional SNPs might be found which were not at all considered by Stranger et al. (2007).

Starting point was the set of SNPs from phase 2 of the HapMap project. In accordance with Stranger et al. (2007) we only considered SNPs with MAF > 0.05, which results in a set of 2698476 SNPs. Filtering identical SNPs yields a subset of 2145627 SNPs, many of which are strongly correlated. In that situation many of the SNPs do not bring substantially new information to the genotype data. To solve this problem in Bogdan et al. (2008a) the notion of 'effective number' of markers was introduced, an idea which can be also found e.g. in Nicodemus et al. (2005). The 'effective number' of markers is used to replace the number of available regressors in the penalty for modifications of BIC. In our real data analysis the 'effective number' of SNPs is calculated based on the clustering algorithm described in Frommlet (2010). This algorithm yields clusters of SNPs which all have pairwise correlation above a chosen threshold C. In accordance with results of Section 3.2 we have chosen $C = 0.7$, which leads to an effective number of 780675 SNPs.

We performed model selection based on mBIC2 (13) with $p = 780675$ and we applied the search algorithm described in Section 3.1. We observed that in some cases the stepwise selection procedure got trapped in local minima for models which are too large. Therefore we added a final step to the search strategy, where we performed an all subset selection over the combined set of SNPs detected by mBIC2 and by Stranger et al. (2007). If this set of SNPs was excessively large (>25) we performed backward selection, and all subset selection only for models including less than 5 SNPs. All cis and trans SNPs detected in this way are presented in Web_Supplement_1.txt. In Web_Supplement_2.txt we show correlations between all SNPs found by mBIC2 and by Stranger et al. (2007). Table 2 shows a comprehensive summary of these results.

As discussed in Section 4.1 SNPs found by model selection are naturally representatives of a number of correlated SNPs (see Web_Supplement_1.txt). On the other hand many SNPs detected by the multiple testing approach from Stranger et al. (2007) are strongly correlated and frequently even have identical genotype for all individuals. To make results comparable we selected representatives of correlated SNPs from Stranger et al. (2007) by applying Tagger (de Bakker et al., 2005), a tag SNP selection algorithm implemented in Haploview (Barrett et al., 2005). In accordance with the discussion in Section 4.1 we used as threshold $|R| = 0.7$ (i.e. $R^2 = 0.49$).

Table 2 provides the number of tag SNPs for cis and for trans associations for the 44 genes with trans association reported in Stranger et al. (2007). Furthermore, we provide the number of cis and trans association detected when using mBIC2, first for models with dummy variables corresponding to different populations, then for models without such dummy variables. We also report the number of matches between Stranger et al. (2007) and our model selection approach, where we define that a match occurs when the absolute correlation between a tag SNP and a SNP detected by mBIC2 is larger than 0.7.

For models considering population structure we report p-values of F Tests on the overall effect of the 3 dummy variables. Taking into account population stratification is important. Without including dummy variables in most genes the number of detected trans SNPs is inflated. Almost all of these additional findings are associated with population structure, which corresponds well with the small p-values observed in column three. For most genes the expression levels vary between populations, and among the huge number of SNPs there will always be some which pick up this variation.

Results are arranged according to the categories presented in the last column. In category A we collect 19 genes where the model with dummy variables found all SNPs from Stranger et al. (2007), in the 12 genes of category B it did not find any of them, and in the 6 genes of category A/B it detected some but not all of them. Category C collects 7 genes for which Stranger et al. (2007) reports an extraordinary large number of SNPs. When we take into account population stratification, then for the majority of genes of category A and A/B our results are quite similar to those of Stranger et al. (2007). In category B there are 7 genes for which Stranger reported 1 or 2 associated SNPs which were not detected using mBIC2. This is not surprising given the fact that we were penalizing for a much larger number of markers.

On the other hand, results for the genes hmm25278-S, hmm26651-S, hmm34610-S and hmm32074-S are very interesting. These genes are located on chromosome 1, 20, 8 and 6 respectively, but their expression levels are strongly

Table 2
Summary statistics for the 44 genes reported in Stranger et al. (2007). Col. 2: Number of tag SNPs representing detections by Stranger et al. (2007) for cis and trans association {original number in curly brackets}. Col. 3 and 4: Number of SNPs detected by mBIC2 as well as number of matches with and without taking into account population structure [number of cis SNPs within box]. Col. 3 also shows *p*-values of F-Test for dummy variables. Col. 5: Categories of genes as described in the main text.

Gene	Stranger		With dummy			No dummy		
	Cis	Trans	SNPs	Match	pval	SNPs	Match	Cat
GI_14277699-S		1	1	1	1.3E−16	5	1	A
GI_15718725-S		1	1	1	3.3E−08	3	1	A
GI_21536317-S		1	2	1	1.9E−05	6		A
GI_22749298-S		{3} 1	1	1	4.6E−05	2	1	A
GI_25952101-I		1	1	1	1.6E−08	2	1	A
GI_34147704-S		{3} 1	1	1	5.6E−13	2	1	A
GI_37545699-S		1	1	1	2.3E−09	3		A
GI_37552052-S		1	1	1	4.4E−20	3	1	A
GI_39841070-S		1	1	1	1.4E−14	3	1	A
GI_41147791-S		{3} 1	[1] 4	1	1.4E−23	3	1	A
GI_42655578-S		1	1	1	2.8E−17	2		A
GI_42656964-S		1	1	1	0.0023	2	1	A
GI_42659691-S		1	1	1	2.8E−08	3	1	A
GI_42662536-S		{15} 1	2	1	1.2E−09	1	1	A
hmm26651-S		{3} 2	5	2	2.3E−06	8	1	A
hmm32074-S		1	11	1	1.4E−42	9	1	A
hmm32535-S		1	1	1	5.2E−22	5	1	A
Hs.292310-S		{2} 1	1	1	0.0002	5	1	A
Hs.514777-S		1	3	1	2.3E−52	7	1	A
GI_18765712-S	{7} 1	1	[1] 0	[1] 0	1.2E−20	[1] 3	[1] 0	A/B
GI_22062109-S	1	1	[1] 0	[1] 0	2.6E−07	[1] 2	[1] 0	A/B
GI_42660576-S		{4} 2	1	1	0.0002	2	1	A/B
hmm25278-S		{3} 2	4	1	9.1E−15	4	1	A/B
hmm34610-S		2	7	1	5.2E−07	4	1	A/B
Hs.517172-S	{11} 2	{2} 1	[1] 2	[1] 1	3.9E−05	[2] 2	[2] 1	A/B
GI_16753224-S		1	1		4.9E−27	5		B
GI_21237760-S		1			3.2E−10	2		B
GI_22325391-S		{2} 1			2.4E−05	2		B
GI_31543145-S		1	[1] 1		0.0015	[1] 2		B
GI_34147394-S		1			2.2E−06	1		B
GI_37552433-S		{2} 1	1		2.2E−18	3		B
GI_38679899-S	{3} 1	1			3.8E−06	2		B
GI_38679979-A		1			0.0003	0		B
GI_40316914-S		1			1.3E−08	3	1	B
GI_42659564-S		{4} 1	2		1.9E−21	4		B
GI_9790904-S		1	1		1.5E−12	1		B
Hs.435267-S		1			7.4E−12	2		B
GI_10864076-S		{21} 5	1	1	1.5E−07	4	1	C
GI_19557676-S	[98] 12	{7} 1	[2] 1	[2] 0	0.01	[2] 1	[2] 0	C
GI_23510353-S		{27} 5	1	1	0.24	2	1	C
GI_33469144-S		{57} 1	1		0.0001	3		C
GI_37537711-S		{17} 5	3	1	1.6E−06	8	2	C
GI_41150880-S		{42} 11	1	1	2.2E−15	10	3	C
GI_42657060-S		{53} 11	3	4	1.0E−10	8	5	C

correlated (pairwise correlation larger than 0.92 for each possible combination). Stranger et al. (2007) reported the following trans SNPs: rs9528181 for all four, rs7318180 for the first three, and rs12860901 for the first two or them. These SNPs are all located on chromosome 13 at positions 113893447, 113835272 and 113901892, respectively, which indicates that this region on chromosome 13 has strong regulatory influence on the four genes under discussion.

For all these genes model selection based on mBIC2 is finding larger models, which are summarized in Table 3. All four models include a SNP on chromosome 13 which represents the SNPs reported in Stranger et al. (2007). Furthermore, all four models include trans SNPs on chromosome 2 and on chromosome 3, though not all of them are located in the same regions of chromosome 2 and chromosome 3, respectively. For hmm25278-S there is one more SNP on chromosome 15 which does not correspond to any of the other detections. Models for the other three genes agree on SNP rs2044109 on chromosome 8, and they all include a SNP on chromosome 1. The models of hmm32074-S and hmm34610-S include further non-corresponding SNPs. In summary, according to our study there is strong evidence that more than one region has regulatory influence on the expression levels of these four genes. Also this example shows that for the future a multivariate approach taking into account the information of correlated traits might be of some interest.

Table 3
Models selected by mBIC2 for the genes hmm25278-S, hmm26651-S, hmm32074-S and hmm34610-S. Each column describes the markers included in the best model of the respective gene. SNP name (first line), chromosome and position (second line) are provided for each selected SNP, with the exception of 5 selected SNPs for gene hmm32074-S, which can be found in Web_Supplement_1.txt.

hmm25278-S		hmm26651-S		hmm32074-S		hmm34610-S	
rs9525262		rs9525181		rs9525262		rs9525262	
13	113891161	13	113893447	13	113891161	13	113891161
rs10048748		rs10490450		rs17386102		rs10490450	
2	165704573	2	33186442	2	17197272	2	33186442
rs10937559		rs6441934		rs1370718		rs6441934	
3	194105335	3	45937806	3	32328135	3	45937806
rs8028606		rs2044109		rs2044109		rs2044109	
15	92857298	8	3074517	8	3074517	8	3074517
		rs17455546		rs2819755		rs17455546	
		1	100637742	1	236089656	1	100637742
				rs13021147		rs3761945	
				2	107939438	1	228773391
				5 more SNPs on		rs17326215	
				Chr. 5, 9, 10, 18, 22		7	24408655

For the genes discussed above the three trans SNPs detected by Stranger et al. (2007) are located very close to each other on the same chromosome. This is actually typical: For all 44 genes, the trans SNPs reported by Stranger et al. (2007) are located within a relatively small region. This holds even for the 7 genes of category C, which are characterized by an untypically large number of SNPs reported in Stranger et al. (2007). These SNPs have a rather complex correlation structure, but their positions are for all cases within less than 400 kb.

If we take for example gene GI_19557676-S, the reported cis SNPs (chromosome 6, between pos. 31105671 and 31439808) and trans SNPs (chromosome 6, between pos. 30045241 and 30049163) are located fairly close to each other. One might think of an extended cis region, and mBIC2 is finding 3 SNPs (2 cis, one trans) which represent the genetic variability within that region. Although the trans SNP rs3823342 (chr. 6, pos. 30021046) found by model selection is not a match according to our definition based on correlation, it indicates the same region. The same is true for gene GI_33469144-S, where SNP rs2996607 on chromosome 10 is in the same region as all the trans SNPs reported by Stranger et al. (2007), though based on the correlation criterion it does not count as a match.

If we look at the genes GI_10864076-S (Chr. 16) and GI_23510353-S (Chr. 19), in both cases mBIC2 found one matching trans SNP which turns out to be strongly correlated with all SNPs reported by Stranger et al. (2007), namely $|R| > 0.49$ for GI_10864076-S and $|R| > 0.48$ for GI_23510353-S. Now interestingly, for genes GI_37537711-S (Chr. 5), GI_41150880-S (Chr. 18) and GI_42657060-S (Chr. 4) the trans SNPs found by Stranger et al. (2007) are all lying exactly in the same region as those of GI_10864076-S and GI_23510353-S (Chr. 6, between position 32500000 and 32800000), and also many trans SNPs are actually shared by these genes. It is clear that this region must have a particularly strong regulatory effect on other genes, that is susceptible to genetic variability. Multiple testing strategies pick up many correlated SNPs reflecting these signals, whereas mBIC2 is detecting a smaller number of SNPs representing that region.

Finally we want to mention several other genes for which additional trans SNPs have been found, namely GI 21536317-S, GI 31543145-S, GI 41147791-S, GI 42659564-S, GI 42662536-S, Hs.514777-S and Hs.517172-S. Perhaps most remarkable among those are GI 41147791-S and GI 31543145-S, where the model selection approach was able to detect a cis SNP which was not detected by multiple testing.

6. Discussion

We have introduced a model selection approach for genome-wide association studies using modifications of BIC which are based on sound theoretical considerations (Bogdan et al., 2008b; Frommlet et al., submitted for publication). Elementary statistical arguments have shown that model selection is preferable to multiple testing strategies based on single-marker analysis, and a comprehensive simulation study confirmed this. Our model selection procedure, using modifications of mBIC and a relatively simple search strategy, had on average considerably larger power than single-marker analysis, while at the same time controlling type I error rates at a lower level. For the ease of presentation we restricted the discussion to linear regression models, though qualitatively similar results will hold for generalized linear models, in particular for logistic regression models in the case of dichotomous traits.

The superior performance of model selection procedures in GWAS has been pointed out by several other studies. Hoggart et al. (2008) developed penalized likelihood criteria on Bayesian ideas which are very similar in spirit to our work. Their criterion is based on the particular choice of some shrinkage priors, whereas our BIC type approach does not require any particular specification of prior distributions. The close relationship between mBIC2 and EBIC, which was applied to GWAS in Zhao and Chen (submitted for publication), was already mentioned in Section 2.2. Model selection based on lasso for

GWAS was studies by Wu et al. (2009) and Kooperberg et al. (2010), while very recently in Li et al. (2010) a version of the Bayesian lasso was applied to GWAS. Furthermore, in He and Lin (2011) a model selection approach based on iterative sure independence screening was suggested. All these papers agree on the fact that model selection is much more powerful to detect causal variants in GWAS than single-marker analysis. The individual benefits of these different approaches are still open to be analyzed in a comparative study. In the future an interesting fully Bayesian approach might be based on the methods described in Kwon et al. (in press).

Apart from the expected result that model selection strategies outperform single-marker analysis in terms of power the most important result we obtained is that under complex models one cannot trust the order of p-values from single-marker models. Test statistics are highly influenced by random small correlations to causal SNPs, leading on the one hand to a large number of false positives, on the other hand to severely reduced power for some important causal mutations. This loss of power might be one aspect of the widely discussed phenomenon of missing heritability in GWAS (for a recent discussion see Manolio et al. (2009)). It is believed that missing heritability might be found in rare SNPs, or that epigenetic effects might play an important role (McCarthy and Hirschhorn, 2008). However, our results indicate that the statistical analysis performed is an important aspect of the problem, and that single-marker analysis is just not really well suited for GWAS analysis.

Finally we performed a real data analysis based on 210 individuals from the HapMap project, where model selection provided some interesting detections not found by the original analysis based on multiple testing. One aim of Stranger et al. (2007) was to detect association with gene expression levels of SNPs lying outside the region of the considered gene (trans regulatory SNPs). 44 genes with trans regulatory SNPs were reported when pooling over all HapMap populations. We were able to increase the number of detected trans regulatory regions in several cases.

In the simulation study model selection performed unambiguously better than multiple testing. In real data analysis, compared to the original analysis, a substantial number of new putative regions of trans association could be found. Still, the effects were not as strong as in the simulation study; the largest model included 11 SNPs, two models were of size 7 and 5, the rest of size 4 or smaller. We believe that this is mainly due to the rather small sample size. To select more complex models one would need studies with a larger number of individuals, for which it is expected that differences between multiple testing and model selection are getting even more pronounced.

To deal with the huge number of potential models we introduced a rather simple search strategy designed for this particular application. Our search strategy served well in the simulation study, but it had some limitations in the real data analysis. The focus of this manuscript was not on search strategies. Modifications of ideas presented in Baierl et al. (2006) in the context of QTL mapping might be useful. Other possible approaches have been discussed in Zhao and Chen (submitted for publication) and Chen and Chen (2009). The exact choice of model search strategies in GWAS is certainly a fruitful topic for further research.

Acknowledgments

This research was funded by the WWTF project MA09-007a. We thank the two anonymous referees for very constructive comments to improve the manuscript.

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.05.005.

References

Abramovich, F., Benjamini, Y., Donoho, D.L., Johnstone, I.M., 2006. Adapting to unknown sparsity by controlling the false discovery rate. Ann. Statist. 34, 584–653.

Baierl, A., Bogdan, M., Frommlet, F., Futschik, A., 2006. On locating multiple interacting quantitative trait loci in intercross designs. Genetics 173, 1693–1703.

Baierl, A., Futschik, A., Bogdan, M., Biecek, P., 2007. Locating multiple interacting quantitative trait loci using robust model selection. CSDA 51, 6423–6434.

de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., Altshuler, D., 2005. Efficiency and power in genetic association studies. Nat. Genet. 37, 1217–1223.

Balding, D.J., 2006. A tutorial on statistical methods for population association studies. Nat. Rev. Genet. 7, 781–791.

Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263–265.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B. 57, 289–300. MR1325392.

Bogdan, M., Frommlet, F., Biecek, P., Cheng, R., Ghosh, J.K., Doerge, R.W., 2008a. Extending the Modified Bayesian Information Criterion (mBIC) to Dense Markers and Multiple Interval Mapping. Biometrics 64, 1162–1169.

Bogdan, M., Zak-Szatkowska, M., Ghosh, J.K., 2008b. Selecting explanatory variables with the modified version of Bayesian Information Criterion. Qua. Reliab. Eng. Int. 24, 627–641.

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

Broman, K.W., Speed, T.P., 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. J. R. Stat. Soc. Ser. B. (Stat. Methodol.) 64 (4), 641–656.

Chen, J., Chen, Z., 2008. Extended Bayesian Information criteria for model selection with large model spaces. Biometrika 95 (3), 759–771.

Chen, J., Chen, Z., 2009. Tournament screening cum EBIC for feature selection with high-dimensional feature spaces. Sci. China Ser. A Math. 52 (6), 1327–1341.

Chen, J., Chen, Z., 2010. Extended BIC for small *n*-large-*P* sparse GLM (submitted for publication) Available at www.stat.nus.edu.sg/~stachenz/ChenChen.pdf.
Chipman, H., George, E.I., McCulloch, R.E., 2001. The practical implementation of Bayesian model selection (with discussion). In: Lahiri, P. (Ed.), Model Selection. IMS, Beachwood, OH, pp. 66–134.
Dudbridge, F., Gusnanto, A., 2008. Estimation of significance thresholds for genomewide association scans. Genet. Epidem. 32, 227–234.
Erhardt, V., Bogdan, M., Czado, C., 2010. Locating multiple interacting quantitative trait loci with the zero-inflated generalized Poisson regression. Stat. Appl. Genet. Mol. Biol. 9 (1), Article 26.
Frommlet, F., 2010. Tag SNP selection based on clustering according to dominant sets found using replicator dynamics. Adv. Data Anal. Classif. 4, 65–83.
Frommlet, F., Chakrabarti, A., Murawska, M., Bogdan, M., 2010. Asymptotic Bayes optimality under sparsity of selection rules for general priors (submitted for publication) currently available at arXiv:1005.4753.
George, E.I., Foster, D.P., 2000. Calibration and empirical Bayes variable selection. Biometrika 87, 731–747.
He, Q., Lin, D., 2011. A variable selection method for genome-wide association studies. Bioinformatics 27 (1), 1–8.
Hirschhorn, J.N., Daly, M.J., 2005. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6 (2), 95–108.
Hoggart, C.J., Whittaker, J.C., De Iorio, M., Balding, D.J., 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLOS Genet. 4 (7), e1000130. doi:10.1371/journal.pgen.1000130.
Kooperberg, C., LeBlanc, M., Obenchain, V., 2010. Risk prediction using genome-wide association studies. Genet. Epidem. 34, 643–652.
Koren, M., Kimmel, G., Ben-Asher, E., Gal, I., Papa, M.Z., Beckmann, J.S., Lancet, D., Shamir, R., Friedman, E., 2006. ATM haplotypes and breast cancer risk in Jewish high-risk women. Br. J. Cancer 94 (10), 1537–1543.
Kwon, D., Landi, M.T., Vannucci, M., Issaq, H.J., Prieto, D., Pfeiffer, R.M., 2011. An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *CSDA*, Accepted Manuscript, Available online 4 May 2011, in press (doi:10.1016/j.csda.2011.04.019).
Lao, O., et al., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Curr. Biol. 18 (16), 1241–1248.
Li, J., Das, K., Fu, G., Li, R., Wu, R., 2010. The bayesian lasso for genome-wide association studies. Bioinformatics 27 (4), 516–523.
Madow, W., 1940. The distribution of quadratic forms in noncentral normal random variables. Ann. Math. Stat. 11, 100–103.
Manolio, T.A., et al., 2009. Finding the missing heritability of complex diseases. Nature 461 (7265), 747–753.
McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., Hirschhorn, J.N., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9 (5), 356–369.
McCarthy, M.I., Hirschhorn, J.N., 2008. Genome-wide association studies: potential next steps on a genetic journey. Hum. Mol. Genet. 17, R156–R165.
Nelson, M.R., et al., 2008. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet. 83, 347–358. Epub 2008 Aug 28.
Nicodemus, K.K., Liu, W., Chase, G.A., Tsai, Y.Y., Fallin, M.D., 2005. Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms. BMC Genet. 6 (suppl. 1).
Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Montgomery, S., Tavari, S., Deloukas, P., Dermitzakis, E.T., 2007. Population genomics of human gene expression. Nature Genet. 39, 1217–1224.
The International HapMap Consortium, , 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–862.
Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K., 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25 (6), 714–721.
Żak, M., Baierl, A., Bogdan, M., Futschik, A., 2007. Locating multiple interacting quantitative trait loci using rank-based model selection. Genetics 176 (3), 1845–1854.
Żak-Szatkowska, M., Bogdan, M., 2011. Modified versions of bayesian information criterion for sparse generalized linear models. *CSDA*. Accepted Manuscript, Available online 1 May 2011, in press (doi:10.1016/j.csda.2011.04.016).
Zhao, J., Chen, Z., 2010. A two-stage penalized logistic regression approach to case-control genome-wide association studies (submitted for publication) Available at www.stat.nus.edu.sg/~stachenz/MS091221PR.pdf.
Ziegler, A., König, I.R., Thompson, J.R., 2008. Biostatistical aspects of genome-wide association studies. Biometrical Journal 50 (1), 8–28.