A model selection approach to genome wide association studies

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

[a]*Department of Statistics, University of Vienna, Austria*
[b]*Institute of Mathematics and Computer Science, Wrocław University of Technology, Poland*

Abstract

For the vast majority of genome wide association studies (GWAS) published so far, statistical analysis was performed by testing markers individually. In this article we present some elementary statistical considerations which clearly show that in case of complex traits the approach based on multiple regression or generalized linear models is preferable to multiple testing. We introduce a model selection approach to GWAS based on modifications of Bayesian Information Criterion (BIC) and develop some simple search strategies to deal with the huge number of potential models. Comprehensive simulations based on real SNP data confirm that model selection has larger power than multiple testing to detect causal SNPs in complex models. On the other hand multiple testing has 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. We show that this behavior is typical in GWAS for complex traits and can be explained by an aggregated influence of many small random sample correlations between genotypes of a SNP under investigation and other causal SNPs. We believe that our findings at least partially explain problems with low power and nonreplicability of results in many real data GWAS. Finally, we discuss the advantages of our model selection approach in the context of real data analysis, where we consider publicly available gene expression data as traits for individuals from the HapMap project.

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

1. Introduction

Within the last 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. [5, 30, 35, 49]. The major goal of GWAS is to detect association between some trait (either quantitative or categorical) and some genetic markers. The most commonly used

type of markers are single nucleotide polymorphisms (SNPs). Current SNP array technology allows 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 [49]). It is common practice in applied papers on GWAS to report single marker tests of SNPs. For a review on statistical tests for case control studies we refer again to [49]. Recommended significance levels to control family wise error are as small as $\alpha = 5 \cdot 10^{-8}$ [23], 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 permutation tests like in [47] or the use of Hidden Markov Models [48].

Most GWAS are performed as case control studies, though recently there has been growing interest in GWAS for quantitative traits (QT) [43]. Statistical analysis performed for QT tends to make use of regression models to correct for covariates like sex or age; however, each SNP is then tested individually at a significance level accounting for multiple testing (see for example [21, 28, 37, 39, 42] etc.).

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 [16] it was noticed that classical model selection criteria like Akaike Information Criterion (AIC, [2]), and even Bayesian Information Criterion (BIC, [44]), tend to select too many markers. Addressing this problem [12] 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 [11] and [15] 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 [3, 4, 11, 15, 24, 51] based on simulation studies good properties of mBIC were documented. Recently some asymptotic optimality properties of mBIC have been shown [10, 26].

In this article we will adopt a model selection approach for GWAS, which will be introduced in Section 2. For the ease of presentation we will restrict our discussion mainly to linear regression models, though qualitatively similar results will hold for generalized linear models. In Section 2.1 we present basic statistical considerations, which clearly demonstrate the advantage of using multiple regression over the search over single markers. In this section we specifically stress the lower power of single marker tests. This results from an inflated residual sum

of squares, which incorporates the influence of all causal genes that are not included in the model. In Section 2.2 we introduce and motivate our particular choice of model selection criteria for the high dimensional multiple regression we are dealing with. Apart from mBIC we consider a second modification of BIC, mBIC2, which according to [26] is closely related to the multiple regression version of the Benjamini-Hochberg procedure [8] for multiple testing. According to [26] mBIC2 has asymptotic optimality properties in a wider range of sparsity parameters than mBIC. The greatest challenge in applying model selection to GWAS is the huge search space of potential models. In Section 2.3 we describe some model search strategies which are particularly suited to the situation where we have a huge number of markers but expect only a small number of them to be strongly associated with the trait.

In Section 3 we perform a simulation study based on actual SNP data. Apart from the expected result that model selection strategies outperform multiple testing procedures the simulation study provides some rather surprising insights. In particular we will see that for multiple testing procedures rather small random correlations between causal SNPs have drastic influence on the order of p-values. This results in a low power to detect some important causal mutations, as well as in a large number of spurious detections. Thus, our results show that single marker tests can lead to many erroneous results, which makes the use of these procedures in the context of GWAS rather questionable.

Finally in Section 4 we reanalyze publicly available gene expression data [45, 46, 47] as quantitative traits for the individuals genotyped in the HapMap project [29]. One aim of [47] 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. Using our model selection approach we are able to increase the number of detected trans regulatory regions in several cases.

2. Methods

We first want to motivate why we advocate a model selection approach. The discussion in this article is in the context of linear regression models, which allows to focus on the principal ideas involved. We expect that similar conclusions as presented here will also hold for generalized linear models, and in particular for logistic regression, which might be applied in case control studies.

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 $1 \leq j^\star_{l_1} < \ldots < j^\star_{l_k} \leq p$, 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 additive model

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

holds. 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 (we consider only models including the intercept). This is a gigantic number taking into account that in GWAS we are usually dealing with $p \approx 10^7$. 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 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

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

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 of the form

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

are considered.

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 non-central F-distribution and power calculations for different effect sizes are rather straight forward (compare results for $j = 1$ in Figure 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

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

and according to the generalization of Cochran's theorem for the noncentral case [33] the model sum of squares and the residual sum of squares are independent with distributions

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

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

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 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 Figure 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

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

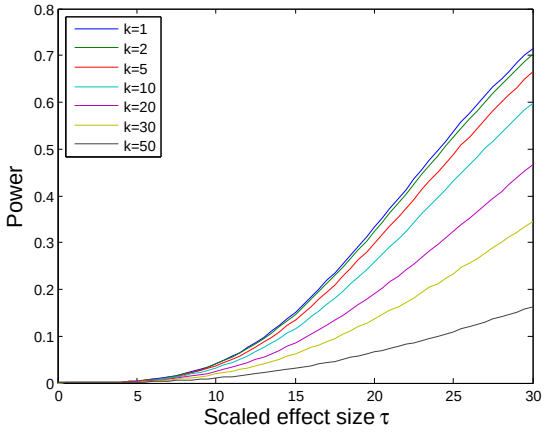

Figure 1: Power to find a causal SNP with single marker tests when model $\mathcal{M}_{\mathbf{j}^*}$ with k effects is correct. In case of $k = 1$ one is testing for the correct model.

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 $MSS_{\mathbf{j}}$. In section 3 we will show that the joint influence of many small random correlations between causal SNPs has a very strong effect on the noncentrality parameters ν_{M_j}. This results in substantial problems with ranking of p-values and leads both to a low power of detection of some of the causal SNPs as well as to a relatively large number of false positives. Also, the general problem of loss of power when testing for single markers under a complex true model remains the same. We thus feel justified to claim that a model selection approach is the favorable alternative to single marker analysis.

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 case of $p > n$. As explained in detail in [15] 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 a modification of BIC was introduced in [12], 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. [20]). 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 [12]).

In 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 [15]. In particular mBIC is controlling the family wise error. In [26] 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 [52] as well as Zero Inflated Generalized Poisson Regression [24].

Recently, in [17] a new modification of BIC, extended Bayesian Information Criterion (EBIC), was proposed. 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 [17] some consistency properties of EBIC are proved under the assumptions that the maximal dimension searched by EBIC is fixed and larger then the true number of effects. In [18] EBIC was further extended to Generalized Linear Models and in [53] 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,

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

which is asymptotically equivalent to EBIC with $\kappa = 0$ when the maximal allowable number of regressors, Q, is of the order $Q = o(p)$.

mBIC2 was developed in [26] as a model selection rule which in the context of multiple regression works similarly to the Benjamini-Hochberg correction for multiple testing. In [26] a thorough discussion is provided how this modification of mBIC relates to a similar criterion suggested by [1] as well as to a modification of the risk inflation criterion RIC, proposed in [27]. Due to the negative extra term

mBIC2 will potentially select larger models than the original mBIC (12). In [26] 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. Search algorithm

An important question when applying a model selection approach to GWAS data is how to deal with the gigantic number of possible models. Some interesting search strategies for the best multiple regression model were recently proposed e.g in [53] and [19]. However, these advanced model selection strategies are rather unfeasible for the 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 popular 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, where more time consuming search procedures are out of question. In the real data analysis of Section 4 the procedure will be amended with a final step, where all subsets of a specified set of SNPs are considered.

3. Simulation study

Simulations are based on SNP data from the population reference sample POPRES [38]. The dataset 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 [32]. 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 i-th 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 i-th individual are not missing. If we find individuals who have exactly the same values as the i-th 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. 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 was 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. 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 considered, 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. We believe that the phenomena discussed further in this paper, which result from a large number of causal SNPs, can play a role in explaining the problem of 'missing heritability' in GWAS [34], a point which we extensively discuss in Section 3.2.

Each simulated data set was analyzed using multiple testing procedures (Bonferroni and Benjamini Hochberg) as well as model selection approaches based on mBIC and mBIC2. 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 correponding 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. [15]). Based on the calculations of [12] 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: Any detected SNP whose correlation to a causal SNP has absolute value larger than a given threshold is counted as a true positive, otherwise as a false positive. We initially used a threshold of $|R| = 0.9$, and based on simulation results decided to report alternatively also results for a threshold $|R| = 0.7$. Here $|R|$ is the maximum absolute value over all correlations with causal SNPs.

For multiple testing procedures it frequently happens 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 definition we estimate the power of detection for each individual causal SNP, as well as the false discovery rate (FDR).

The graphs in Figure 2 show the observed FDR values for all 1000 replicates based on the thresholds $|R| = 0.9$ and $|R| = 0.7$ respectively. 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 much larger power. The false discovery rate of BH at level 0.05 is apparently larger than FDR of mBIC2 with the constant $d = -2\log 4$, though in absolute terms the number of false positives is often smaller for BH. The choice of the threshold for a "false positive" has a strong effect on FDR. For all procedures FDR is

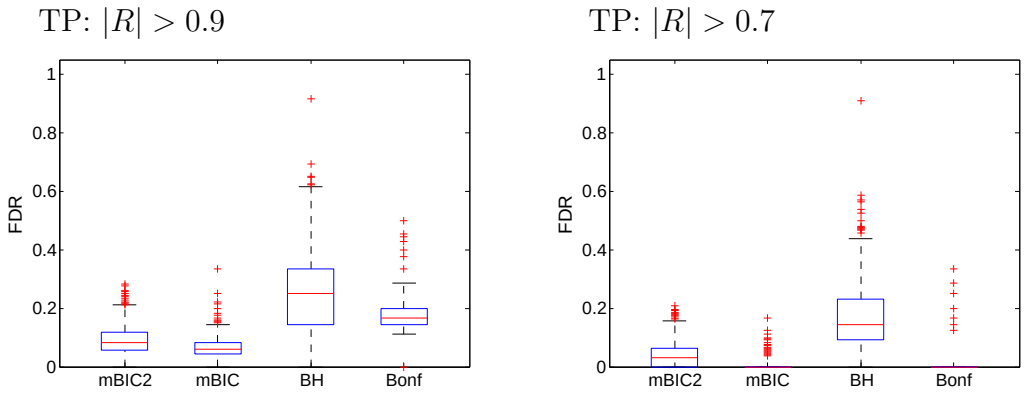

Figure 2: Observed FDR of the 4 different selection procedures. A detected SNP was classified as a true positive when its maximal correlation to a causal SNP was larger than 0.9 in the first graph, and larger than 0.7 in the second graph.

significantly reduced when using the more liberal criterion $|R| = 0.7$. In particular the Bonferroni procedure detects in that case hardly any false positives. The effect of the threshold on the number of false positives and on power is discussed in more detail in Section 3.1.

The graphs in Figure 3 provide the estimated power to detect each of the 40 causal SNPs at the threshold levels $|R| = 0.9$ and $|R| = 0.7$ respectively. The x-axis shows the individual heritability (15) of each SNP. It is evident that mBIC2 has the largest power among the 4 different procedures. Also, as expected, mBIC has in most cases larger power than the two multiple testing procedures.

Most remarkable is the dependence of power on the individual heritability. As expected there is a general trend that larger individual heritability yields larger power, but Figure 3 shows that there is also a huge amount of irregularity. For mBIC2 in general the association between individual heritability and power is quite strong. Still, when using threshold $|R| = 0.9$, there are several SNPs with relatively small power. The most striking example is SNP A-1912140, with a large heritability of 0.03 and power of approximately only 33%. This specific case will be explained in detail in Section 3.1.

For the multiple testing procedures the behavior is much more erratic. The order of SNPs with respect to power is entirely different from the order in terms of individual heritability. For example, the SNP with the largest heritability, easily detected by mBIC2 and mBIC, is completely missed by the procedures based on individual tests. On the other hand, some substantially "weaker" SNPs, are detected with a power exceeding 50%. The key to understand this phenomenon is the influence of correlation between causal SNPs on the noncentrality parameter of the model sum of squares, as discussed in Section 3.2. We will see that this

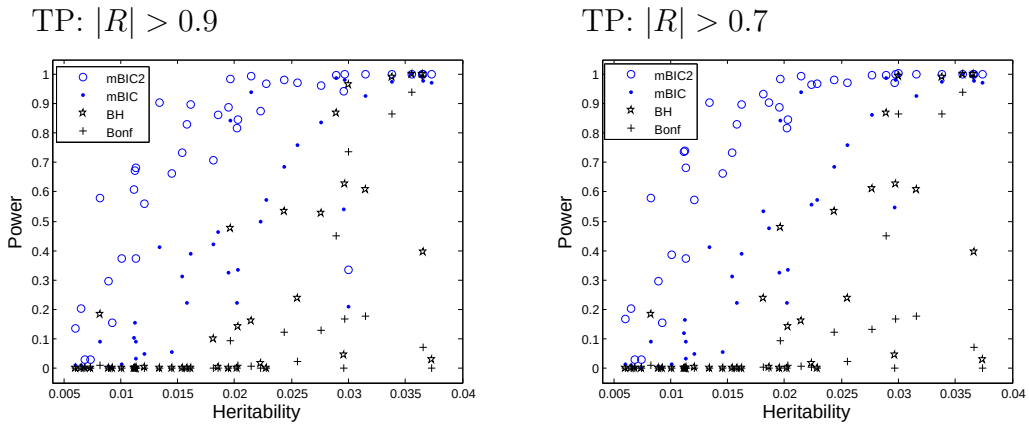

Figure 3: Power of the 4 different selection procedures.

has not only a negative effect on detecting causal SNPs, but it also gives rise to numerous detections of false positive SNPs which have no relation at all to the quantitative trait. This result we consider to be the most important outcome of this simulation study.

3.1. Dependence of TP and FP on $|R|$ - thresholds

Table 1 shows how often certain SNPs occur as false positives based on the threshold $|R| = 0.9$ for mBIC2 and for BH. The first significant finding is that SNP A-2299101, detected in 670 simulation runs, has a correlation of 0.8958 with causal SNP A-1912140. This explains the low power of 33% to detect causal SNP A-1912140, 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 column of Table 1 we observe that all SNPs detected by mBIC2 in more than 5 simulation runs have $|R| > 0.5$. 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 suitable 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. 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 as false positives only once (1076), twice (55) or three times (12). These detections, which might be classified as actual false positives, are usually not correlated to any causal SNP. Based on a model selection approach one would expect to detect some false positives of this nature, and their frequency

Table 1: Thirty most frequent false positive SNPs (based on $|R| < 0.9$) under mBIC2 and under BH. First the SNP name, then the frequency in how many simulation runs the SNP was detected as a false positive, and finally the absolute correlation $|R|$ to the closest causal SNP.

mBIC2			BH		
SNP	freq	corr	SNP	freq	corr
SNP_A-2299101	670	0.8958	SNP_A-2299101	990	0.8958
SNP_A-2034806	133	0.8416	SNP_A-2181789	354	0.2628
SNP_A-2170607	92	0.7728	SNP_A-1804206	136	0.7187
SNP_A-4270622	63	0.6255	SNP_A-1839674	132	0.1137
SNP_A-2266375	56	0.8372	SNP_A-1839540	128	0.5742
SNP_A-1790281	55	0.7683	SNP_A-2251903	87	0.1116
SNP_A-2048646	47	0.8311	SNP_A-2034806	61	0.8416
SNP_A-4201549	43	0.8351	SNP_A-4236404	58	0.7087
SNP_A-1804206	35	0.7187	SNP_A-1818215	56	0.1259
SNP_A-2101072	33	0.7659	SNP_A-4201549	56	0.8351
SNP_A-2091172	31	0.7230	SNP_A-2167803	52	0.0970
SNP_A-2299237	24	0.8954	SNP_A-2048646	50	0.8311
SNP_A-4231385	23	0.8162	SNP_A-1922491	50	0.7145
SNP_A-2267857	21	0.7871	SNP_A-4291099	40	0.4563
SNP_A-2198243	18	0.8632	SNP_A-1894129	39	0.5479
SNP_A-2293694	16	0.5097	SNP_A-4217508	37	0.1007
SNP_A-2006296	15	0.7277	SNP_A-1810532	34	0.1059
SNP_A-2119492	14	0.5207	SNP_A-2241893	32	0.1373
SNP_A-1839540	12	0.5742	SNP_A-2032742	32	0.1024
SNP_A-4266983	11	0.5207	SNP_A-1804069	24	0.0896
SNP_A-1961183	10	0.8573	SNP_A-1804341	19	0.0994
SNP_A-4241095	8	0.5421	SNP_A-2120788	17	0.8978
SNP_A-1894737	8	0.6876	SNP_A-2213672	16	0.1162
SNP_A-1784603	6	0.7480	SNP_A-2171843	16	0.0997
SNP_A-2308622	6	0.7776	SNP_A-2163734	14	0.1123
SNP_A-1957857	6	0.6638	SNP_A-2294489	13	0.1109
SNP_A-4224627	5	0.1008	SNP_A-4254512	13	0.0875
SNP_A-1965812	5	0.7632	SNP_A-1865448	13	0.1115
SNP_A-1835435	5	0.1284	SNP_A-2051237	12	0.3905
SNP_A-1976469	5	0.3733	SNP_A-1816015	12	0.5199

is controlled according to the theory of mBIC2.

3.2. Multiple testing procedures and heritability

The second column of Table 1 shows the thirty most frequent false positive SNPs under BH. There are several SNPs which coincide with SNPs from the first column. Practically all other SNPs which have not been detected by mBIC2 as false positives have a striking characteristic: They are not strongly correlated to any causal SNP. The most prominent example is SNP A-2181789, which has been detected 354 times as a false positive, but is only correlated with $|R| = 0.2628$ to the closest causal SNP.

We will now provide the explanation for this seemingly implausible result, and will also explain the erratic behavior of power in terms of individual heritability observed in Figure 3. Remember that F-tests of single effect models involve non-centrality 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 case of orthogonal design matrix, $\nu_{M,j}$ is proportional to the individual heritability. However, in the general case the non-centrality 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 Figure 4 shows that small pairwise correlations with other causal SNPs explain the erratic behavior of the multiple testing procedures. When we plot the power against the square root of the non-centrality parameter $\nu_{M,j}$ we observe 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 Figure 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

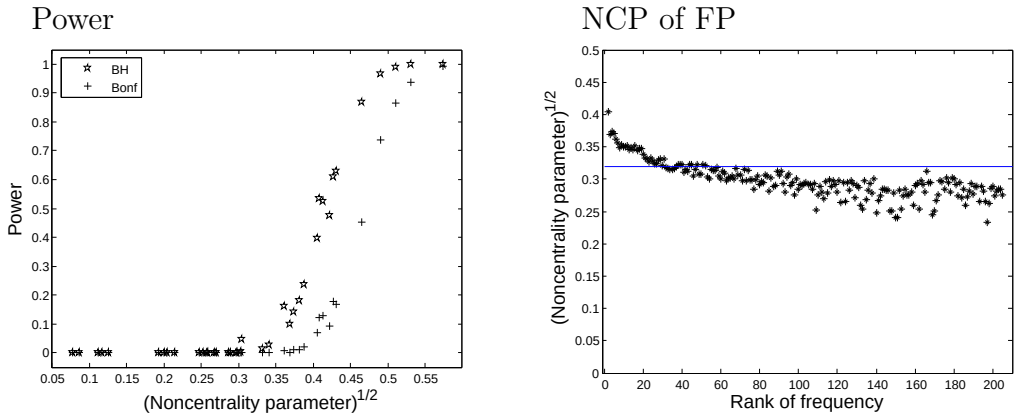

Figure 4: First plot: Power of the multiple testing procedures. Instead of the individual heritability we plot against the square root of the noncentrality parameter. Second plot: Square root of the noncentrality parameter of correlations for all false positives occurring under BH (at level $|R| = 0.9$). On the x-axis SNPs are ordered according to their frequency of detection in the 1000 simulation runs. The initial first stars from the left correspond to the SNPs listed in the second column of Table 1.

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 30 SNPs listed in the second column of Table 1 all have $\sqrt{\nu_{M,j}} > 0.32$. This is just the onset of the sigmoid function observed in the first plot of Figure 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 multiple testing procedures have 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 multiple testing procedures in GWAS with complex traits.

4. Real data analysis

In a series of papers Stranger et al. [45, 46, 47] have analyzed the association between SNPs from the HapMap project [29] and gene expression data. In particular in [47] they considered 270 individuals from four populations, namely 30 Caucasian trios of northern and western European background (CEU), 30 Yoruba trios from Ibaden, Nigeria (YRI), 45 unrelated individuals from Beijing, China (CHB) and 45 unrelated individuals from Tokyo, Japan (JPT). The major objective of [47] 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 [47] 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 [47] used a procedure based on conditional permutations, which was originally described in [31].

We want to compare our model selection approach in particular with the analysis of [47] 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 [47] 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 [47] only some 25000 candidate SNPs were considered as putatively functional SNPs. Unfortunately these SNPs were not clearly specified (the corresponding link in the supplementary material is not providing the relevant list of SNPs), and we decided to search over all available SNPs. As a consequence mBIC2 will use a much larger penalty for multiple testing, on the other hand functional SNPs might be found which were not at all considered by [47].

Starting point was the set of SNPs from phase 2 of the HapMap project. In accordance with [47] 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 [11] the notion of 'effective number' of markers was introduced, an idea which can be also found e.g. in [40]. 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 [25]. This algorithm yields clusters of SNPs which all have pairwise correlation above a chosen threshold C. In accordance with results of Section 3 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 2.3. 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 [47]. 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. Table 2 shows a comprehensive summary of these results.

Table 2: Summary statistics for the 44 genes reported in [47]. Col. 2: Number of tag SNPs representing detections by [47] 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.

	Stranger		with Dummy			no Dummy		
Gene	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

As discussed in Section 3.1 SNPs found by model selection are naturally representatives of a number of correlated SNPs. On the other hand many SNPs detected by the multiple testing approach from [47] are strongly correlated and frequently even have identical genotype for all individuals. To make results comparable we selected representatives of correlated SNPs from [47] by applying Tagger [6], a tag SNP selection algorithm implemented in Haploview [7]. In accordance with the discussion in Section 3.1 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 [47]. 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 [47] 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 [47], 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 [47] 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 [47]. 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 correlated (pairwise correlation larger than 0.92 for each possible combinations). [47] 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.

Table 3: Models selected for the genes hmm25278-S, hmm26651-S, hmm32074-S and hmm34610-S. SNP name (first line), chromosome and position (second line) for each selected SNP.

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 all these genes model selection is finding larger models, which are summarized in Table 3. All four models include a SNP on chromosome 13 which represents the detection of [47]. Furthermore all four models include trans SNPs on chromosome 2 and on chromosome 3, though not all of them are located in proximity. Apart from that there is a certain amount of ambiguity. For hmm25278-S there is one more SNP which does not correspond to any of the other detections. Models for the other three genes agree on SNP rs2044109, 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. Also this example shows that for the future a multivariate approach taking into account the information of correlated traits might be of some interest.

For the genes discussed above the three trans SNPs detected by [47] are located very close to each other on the same chromosome. This is actually typical: For all 44 genes, the reported trans SNPs 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 [47]. 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 [47], 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 [47], 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 [47] 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.

5. Discussion

We have introduced a model selection approach for genome wide association studies using modifications of BIC which are based on sound theoretical considerations [15, 26]. Elementary statistical arguments have shown that a model selection approach is preferable to multiple testing strategies, and a comprehensive simulation study confirmed this. 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.

Perhaps 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. This loss of power might be one aspect of the widely discussed phenomenon of missing heritability in GWAS (for a recent discussion see [34]). It is believed that missing heritability might be found in rare SNPs, or that epigenetic effects might play an important role [36]. However, our results indicate

that the statistical analysis performed is an important aspect of the problem, and that multiple testing strategies are just not really well suited for GWAS analysis.

In our 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. Then 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 [3] in the context of QTL mapping might be useful. Other possible approaches have been discussed in [53, 19]. The exact choice of model search strategies in GWAS is certainly a fruitful topic for further research.

Acknowlegments:

This research was funded by the WWTF project MA09-007.

References

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

[2] Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control 19*, 716–723.

[3] Baierl, A., Bogdan, M., Frommlet, F., Futschik, A. (2006) On Locating Multiple Interacting Quantitative Trait Loci in Intercross Designs. Genetics. **173**: 1693-1703.

[4] Baierl, A., A. Futschik, M. Bogdan, and P. Biecek (2007). Locating multiple interacting quantitative trait loci using robust model selection. *Computational Statistics and Data Analysis 51*, 6423–6434.

[5] Balding, D. J., (2006) A tutorial on statistical methods for population association studies Nat. Rev. Gen. **7**:781–791.

[6] de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., and Altshuler, D. (2005) Efficiency and power in genetic association studies. Nat Genet. **37**, 1217–23.

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

[8] Benjamini, Y. and Hochberg, Y., (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.* **57** 289–300. MR1325392

[9] Bogdan, M., Chakrabati, Frommlet, F., A. and Ghosh, J.K. (2010) Bayes oracle and the asymptotic optimality of the multiple testing procedures under sparsity. *To appear*, currently available at arXiv:1002.3501

[10] Bogdan, M., Chakrabarti A., Ghosh, J.K. (2008). Optimal rules for multiple testing and sparse multiple regression, Technical Report I-18/08/P-003, Institute of Mathematics and Computer Science, Wrocław University of Technology, 2008.

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

[12] Bogdan, M., Ghosh, J. K., and Doerge, R. W. (2004). Modifying the schwarz bayesian information criterion to locate multiple interacting quantitive trait loci. *Genetics*, **167**, 989–999.

[13] Bogdan, M., Ghosh, J. K., Ochman, A. and Tokdar S.T. (2007) On the Empirical Bayes approach to the problem of multiple testing. *Quality and Reliability Engineering International*, **23**, 727–739.

[14] Bogdan, M., Ghosh, J. K. and Tokdar S. T. (2008) A comparison of the Simes-Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. IMS Collections, **Vol.1**, Beyond Parametrics in Interdisciplinary Research: Fetschrift in Honor of Professor Pranab K. Sen, edited by N. Balakrishnan, Edsel Peña and Mervyn J. Silvapulle, pp. 211–230, Beachwood Ohio.

[15] Bogdan, M., Żak-Szatkowska, M., Ghosh, J.K. Selecting explanatory variables with the modified version of Bayesian Information Criterion, *Quality and Reliability Engineering International*, **24**: 627–641, 2008.

[16] Broman, K.W., Speed, T.P. (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** (4): 641–656.

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

[18] Chen, J. and Z. Chen (2010). Extended BIC for small n-large-P sparse GLM. submitted, available at *www.stat.nus.edu.sg/∼stachenz*/ChenChen.pdf.

[19] Chen, J. and Z. Chen (2010). Tournament screening cum EBIC for feature selection with high-dimensional feature spaces *Science in China Series A: Mathematics* **52** (6): 1327–1341.

[20] Chipman, H., George, E.I. and McCulloch, R.E. (2001) The practical implementation of Bayesian model selection (with discussion). In *Model Selection* (P. Lahiri, ed.) 66–134. IMS, Beachwood, OH.

[21] Cho Y.S., et al. (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet.* 41(5):527–534.

[22] Colditz G.A. and Hankinson S.E. (2005) The Nurses' Health Study: lifestyle and health among women. *Nature Reviews Cancer* 5, 388–396

[23] Dudbridge, F., Gusnanto, A. (2008) Estimation of Significance Thresholds for Genomewide Association Scans *Genet. Epid.* **32**: 227–234

[24] Erhardt, V., M. Bogdan and C. Czado (2010). Locating multiple interacting quantitative trait loci with the zero-inflated generalized Poisson regression, *Statistical Applications in Genetics and Molecular Biology*, **Vol 9 : Iss. 1**, Article 26.

[25] Frommlet, F. (2010). Tag SNP selection based on clustering according to dominant sets found using replicator dynamics, *Adv Data Anal Classif* **4**: 65–83

[26] Frommlet, F., Bogdan, M. and Chakrabarti, A. (2010). Asymptotic Bayes optimality under sparsity of selection rules for general priors. *Technical report*, arXiv:1005.4753

[27] George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika.* **87**: 731–747.

[28] Ganesh S.K., et al. (2009) Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat Genet.* **41(11)** 1191–1198.

[29] The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature **449**, 851–862.

[30] Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. **6**(2):95–108.

[31] Koren M, Kimmel G, Ben-Asher E, Gal I, Papa MZ, Beckmann JS, 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.

[32] Lao O., et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Curr Biol.* **18**(16):1241–1248.

[33] Madow, W. (1940) The distribution of quadratic forms in noncentral normal random variables. *Ann. Math. Stat.* **11**, 100–103).

[34] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. (2009) Finding the missing heritability of complex diseases. *Nature.* **461**(7265):747–753.

[35] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* **9**(5):356–369.

[36] McCarthy, M.I. and Hirschhorn, J.N. (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* **17**, R156–R165).

[37] Meisinger C., et al. (2009) A genome-wide association study identifies three loci associated with mean platelet volume. *Am J Hum Genet.* **84** 66–71.

[38] Nelson M.R., et al. (2008) *Am J Hum Genet.* **83** 347–58. Epub 2008 Aug 28.

[39] Newton-Cheh C, et al. (2009) Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet.* [Epub ahead of print]

[40] 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).

[41] Ouwehand W.H. (2009) The discovery of genes implicated in myocardial infarction. *J Thromb Haemost* **7** Suppl 1:305–307.

[42] Potkin S.G., Guffanti G., Lakatos A., Turner J.A., Kruggel F., Fallon J.H., Saykin A.J., Orro A., Lupoli S., Salvi E., Weiner M., Macciardi F. (2009) Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One* **4(8)**: e6501

[43] Potkin,S.G., Turner, J.A., Guffanti,G., Lakatos, A., Torri, F., Keator, D.B., and Macciardi, F. (2009) Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: Methodological considerations, *Cognitive Neuropsychiatry* **14**: (4/5), 391–418

[44] Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

[45] Stranger, B.E, M.S. Forrest, A.G. Clark, M.J. Minichiello, S. Deutsch, R. Lyle, S. Hunt, B. Kahl, S.E. Antonarakis, S. Tavare, P. Deloukas, E.T. Dermitzakis. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics* **1**:e78.

[46] Stranger, B.E., M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, R. Redon, C.P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S.W. Scherer, S. Tavar, P. Deloukas, M.E. Hurles, E.T. Dermitzakis. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848–853.

[47] Stranger, B.E., A.C. Nica, M.S. Forrest, A. Dimas, C.P. Bird, C. Beazley, C.E. Ingle, M. Dunning, P. Flicek, S. Montgomery, S. Tavar, P. Deloukas, E.T. Dermitzakis. (2007). Population genomics of human gene expression. *Nature Genetics* **39**: 1217–1224.

[48] Wei, Z., Sun, W., Wang, K. and Hakonarson, H. (2009) Multiple testing in genome-wide association studies via hidden Markov models, *Bioinformatics* **25**:(21), 2802–2808

[49] Ziegler, A., König, I. R., and Thompson, J.R. (2008) Biostatistical Aspects of Genome-Wide Association Studies, *Biometrical Journal* **50**:1, 8–28

[50] Zhang, C.L., Qi, D.J., Hunter, J.B., Meigs, J.E., Manson, J.E., vna Dam, R.M., Hu, F.B. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene and the risk of type 2 diabetes in large cohorts of U.S. women and men. *Diabetes* **55**: 2645–2648.

[51] Żak, M., A. Baierl, M. Bogdan, and A. Futschik (2007). Locating multiple interacting quantitative trait loci using rank-based model selection. *Genetics 176*(3), 1845–1854.

[52] Żak-Szatkowska M. and M. Bogdan (2010). Applying generalized linear models for identifying important factors in large data bases. *Technical Report I-18/2010/P-001*. Institute of Mathematics and Computer Science, Wroclaw University of Technology, *www.im.pwr.wroc.pl/∼mbogdan/Preprints*.

[53] Zhao, J. and Z. Chen (2010). A two-stage penalized logistic regression approach to case-control genome-wide association studies. submitted, available at *www.stat.nus.edu.sg/∼stachenz/MS091221PR.pdf*.