PLOS | ONE

Analyzing Genome-Wide Association Studies with an FDR Controlling Modification of the Bayesian Information Criterion

Erich Dolejsi[1], **Bernhard Bodenstorfer**[2], **Florian Frommlet**[1]*

[1] Center for Medical Statistics, Informatics, and Intelligent Systems/Section of Medical Statistics, Medical University Vienna, Vienna, Austria, [2] cogiti e.U., Karlstift, Austria

Abstract

The prevailing method of analyzing GWAS data is still to test each marker individually, although from a statistical point of view it is quite obvious that in case of complex traits such single marker tests are not ideal. Recently several model selection approaches for GWAS have been suggested, most of them based on LASSO-type procedures. Here we will discuss an alternative model selection approach which is based on a modification of the Bayesian Information Criterion (mBIC2) which was previously shown to have certain asymptotic optimality properties in terms of minimizing the misclassification error. Heuristic search strategies are introduced which attempt to find the model which minimizes mBIC2, and which are efficient enough to allow the analysis of GWAS data. Our approach is implemented in a software package called MOSGWA. Its performance in case control GWAS is compared with the two algorithms HLASSO and d-GWASelect, as well as with single marker tests, where we performed a simulation study based on real SNP data from the POPRES sample. Our results show that MOSGWA performs slightly better than HLASSO, where specifically for more complex models MOSGWA is more powerful with only a slight increase in Type I error. On the other hand according to our simulations GWASelect does not at all control the type I error when used to automatically determine the number of important SNPs. We also reanalyze the GWAS data from the Wellcome Trust Case-Control Consortium and compare the findings of the different procedures, where MOSGWA detects for complex diseases a number of interesting SNPs which are not found by other methods.

Citation: Dolejsi E, Bodenstorfer B, Frommlet F (2014) Analyzing Genome-Wide Association Studies with an FDR Controlling Modification of the Bayesian Information Criterion. PLoS ONE 9(7): e103322. doi:10.1371/journal.pone.0103322

Editor: María Mar Abad-Grau, University of Granada - Q1818002F, Spain

Received April 22, 2014; **Accepted** July 1, 2014; **Published** July 25, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. The data set used for simulations are available from dbGaP through dbGaP accession number phs000145.v1.p1 at www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id = phs000145.v1.p1. The WTCCC data are available at http://www.wtccc.org.uk/.

Funding: This research has been funded by the Vienna Science and Technology Fund (http://www.wwtf.at/) through project MA09-007a. All three authors received funding. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The co-author Bernhard Bodenstorfer runs the company cogiti e.U., which provides information technology services such as programming, consultancy and technology management. However, this does not alter the authors' adherence to PLOS ONE policies on sharing data and materials. In particular, the program developed has been openly published on sourceforge.net under the GNU GPL version 3 open license for the scientific community to freely use, as the authors have indicated in their submitted paper.

* Email: Florian.Frommlet@meduniwien.ac.at

Introduction

Recently there has been growing interest in model selection approaches to GWAS analysis. Although it is still common practice in published GWAS to perform statistical analysis for each SNP individually, there is increasing awareness that this kind of single marker analysis has certain deficiencies in case of complex traits. Several authors have commented that marginal tests will suffer from lack of power to detect SNPs because the effect of other causal SNPs remains unaccounted for [23,29]. It has been argued that this shortcoming of single marker tests might play a significant role in the widely discussed phenomenon of "missing heritability" in GWAS [43].

A slightly more sophisticated and less known problem is that single marker tests have serious difficulties to rank important SNPs correctly [23]. This is obvious for SNPs which are not directly associated with a trait, but which have an important effect conditional on the presence of other SNPs. However, even in case of SNPs with marginal effects it turns out that due to small sample correlations some important SNPs might have rather small probability to be detected, whereas other SNPs which are not associated at all with the trait might be selected with large probability. This result puts in question the common practice to report those SNPs in GWAS which have lowest ranking marginal p-values.

Given these deficiencies of single marker tests one can expect that the use of multi marker models to analyze GWAS will become more and more important. Multiple linear regression models for quantitative traits and logistic regression models for case control studies have a long history in genetic association studies. To facilitate their use for GWAS there is a strong demand of two things: A thorough theoretical understanding of different model selection strategies in high dimensions to find the regression model which includes important SNPs, as well as the availability of software packages which make modern statistical methodology applicable to GWAS analysis.

Concerning the theory of high dimensional data analysis the last two decades have seen a large number of innovations. One milestone was the development of LASSO [39], which paved the way for a large number of other new approaches to model selection. Bühlmann and van de Geer [12] give a comprehensive presentation of the theoretical foundations of LASSO and its many extensions like adaptive LASSO, group lasso or the elastic net. In the context of GWAS several algorithms have been implemented based on LASSO or one of its extensions [26,30,42].

From a Bayesian perspective the LASSO is equivalent to model selection with a double exponential (DE) distribution as shrinkage prior. Among the first software packages which allowed to perform multi marker analysis of GWAS was HLASSO [29], which uses not only DE priors, but alternatively considers normal exponential Gaussian (NEG) priors. The NEG distribution is more pointed than DE at 0, resulting in the selection of potentially smaller models. More recently a Bayesian version of the LASSO was introduced for GWAS analysis [31].

The LASSO itself was originally developed for model selection problems of moderate size, whereas in GWAS one usually is confronted with up to a million SNPs or more. For such ultra-high dimensional problems, Fan et al. [17] suggested sure independence screening (SIS) as a convenient way of dimension reduction. In case of regression models SIS is nothing else but preselecting a certain number of markers based on marginal tests. After SIS more refined methods like LASSO or SCAD can be applied to select a model. Using tests conditional on that selected model over all remaining markers one can apply another SIS step. Iterating SIS and refined model selection gives the procedure called ISIS.

A startlingly simple but computationally intensive method to improve the performance of model selection procedures in high dimensions is stability selection [33], where random subsamples of the data are drawn repeatedly and a given model selection procedure like LASSO is performed on each of these subsamples. The final model is then obtained by considering those regressors which have appeared consistently over the repeated samples.

One of the more prominent model selection packages for GWAS is GWASelect [26], which combines ISIS with stability selection based on 50 random subsamples, where the refined model selection procedure in ISIS is LASSO. GWASelect itself uses a prespecified size of the model, but there exists a 'dynamic' version d-GWASelect which uses cross-validation to fit the LASSO parameter and which determines the number of selected SNPs using stability selection. Thus in this article GWASelect actually refers to d-GWASelect, the algorithm which allows to determine the number of interesting SNPs.

An alternative approach to model selection in high dimensions which is currently gaining popularity is based on information criteria. Among the large number of SNPs genotyped in GWAS one expects only a moderate number of SNPs to have a strong effect. In such a sparse setting classical information criteria like Akaike's AIC or Schwarz's Bayesian information criterion (BIC) [38] have been shown to select too large models [11]. Consequently, Bogdan et al. [8] introduced a modification of BIC called mBIC which is designed to control the family wise error rate (FWER) of selected markers in sparse regression [9]. A rather similar criterion called EBIC was presented by Chen and Chen [14], where consistency results for EBIC under sparsity were shown even in case when the number of markers is growing faster than the number of observations.

Here we will focus on mBIC2, a modification of BIC which has the property of controlling the false discovery rate (FDR). The false discovery rate is the expected proportion of incorrectly rejected null hypotheses, and was introduced by Benjamini and Hochberg [6] as a measure of type I error control in multiple testing which is less stringent than the family-wise error rate. Frommlet et al. [21] introduced the criterion mBIC2 to control the FDR in a model selection context. Further background information on mBIC2 is given by Frommlet et al. [23], where extensive GWAS simulations of quantitative traits based on real SNP data show that mBIC2 is considerably more powerful to detect causal SNPs than mBIC, while controlling the FDR at a fixed level.

In this article we will focus on case control studies, where the model selection task is performed using logistic regression models. Modifications of BIC for generalized linear models were studied already in the context of QTL mapping [45]. However, for GWAS the task of minimizing the criterion over all possible models is much more challenging than in QTL mapping due to the much larger number of genetic markers involved. In case of logistic regression computing maximum likelihood estimates for each model becomes much more time consuming than for quantitative traits, and therefore one needs to develop rather involved search strategies trying to find models which minimize mBIC2. The resulting algorithm is implemented in the software package MOSGWA, and we compare its performance with single marker tests and with other variable selection methods, in particular with HLASSO [29] and GWASelect [26]. The main reason for this choice is that He and Lin [26] performed already a comparison with several other methods for GWAS analysis, where HLASSO and GWASelect gave the most convincing results. We will report results from an extensive simulation study to compare the performance of the different procedures, and we will reanalyze the GWAS data from the Wellcome Trust Case-Control Consortium (WTCCC) [41].

Methods

Selection criterion

Before describing the algorithmic details of MOSGWA we will formally introduce the selection criterion mBIC2. Consider a GWAS based on p SNPs and n individuals. For a given model including k_M SNPs our model selection criterion is of the form

$$mBIC2 = -2\log L_M^* + k_M \log(np^2/4) - 2\log(k_M!)\ . \qquad (1)$$

Each set of SNPs forms a potential model M, and MOSGWA tries to find that model which minimizes mBIC2. Here L_M^* is the Firth corrected maximum likelihood, which will be discussed in more detail below.

The penalty of the mBIC2 criterion was introduced by Frommlet et al. [21], where its derivation was based on ideas of Abramovich et al. [1]. In particular mBIC2 is closely related to the Benjamini-Hochberg multiple testing procedure [6], and it controls the false discovery rate of detected SNPs. In the context of linear regression certain asymptotic optimality properties of mBIC2 were shown [7,21]. Roughly speaking selection based on mBIC2 minimizes the misclassification error when both p and n are large, while the number of regressors of the correct model is relatively small.

An extensive motivation of mBIC2 can be found in Frommlet et al. [23], where the criterion is applied for linear regression models to analyze GWAS with quantitative traits. In contrast we will focus here on case-control studies, and just like HLASSO [26] and GWASelect [29] we make use of logistic regression to model the disease risk of SNPs. To this end let Y_i denote the disease status of individual $i \in \{1, \ldots, n\}$ ($Y_i = 1$ for a case, $Y_i = 0$ for a control), and let x_{ij} denote the genotype of SNP $j \in \{1, \ldots, p\}$ for individual i. If a model M includes the SNPs $j_1, \ldots, j_k$ then the corresponding logistic regression model can be written as

$$\pi_i := P(Y_i = 1|M,\theta) = \frac{\exp(\beta_0 + \sum_{r=1}^{k} \beta_r x_{ij_r})}{1 + \exp(\beta_0 + \sum_{r=1}^{k} \beta_r x_{ij_r})} , \qquad (2)$$

with the parameter vector $\theta = (\beta_0, \ldots, \beta_k)^T$. These $k+1$ parameters can be routinely estimated by maximizing the corresponding likelihood $L_M(\theta)$, although occasionally the well known problem of separation may occur, where some parameter estimates tend towards infinity [2]. In classical statistical applications, where $p \ll n$, separation typically arises only in case of small sample sizes. For GWAS the sample size is usually very large, but the number of potential regressors is even several orders larger, which results in many combinations of SNPs for which separation occurs. Heinze and Schemper [27] suggested to overcome the problem of separation using a bias corrected version of logistic regression which was originally proposed by Firth [18]. The likelihood of the logistic regression model is multiplied with the corresponding Jeffreys prior, which is just the square root of the determinant of the Fisher information matrix $I(\theta)$. Thus the Firth corrected maximum likelihood from equation (1) is given by

$$L_M^* = \max_\theta L_M(\theta)\sqrt{|I(\theta)|} , \qquad (3)$$

and explicit formulas are given for example by Heinze and Schemper [27]. The Firth-corrected log-likelihood $\log L_M^*$ includes the penalty term $\log\sqrt{|I(\theta)|}$, which guarantees that parameter estimates cannot get excessively large. Note that LASSO based procedures like GWASelect do not run into difficulties with separation because the L^1-penalty yields automatically a shrinkage of parameters.

Search strategy

Having defined the model selection criterion (1) the main task is to find the model which minimizes mBIC2. The resulting problem is an extremely challenging mixed integer program, for which one can only attempt to develop heuristic methods which yield an approximate solution. The search algorithm of the software package MOSGWA repeatedly makes use of a strategy called fast stepwise search (FSS).

The aim of FSS is, starting from some initial model, to perform a search heuristic which finds a model with a smaller value of a given selection criterion. The final call of FSS is performed with the target criterion mBIC2, but within the search it is valuable to work with less stringent criteria to avoid getting stuck in local minima corresponding to models which are missing some of the causal SNPs. Specifically we consider the milder criterion

$$mBIC_{60} := -2\log L_M^* + k_M \log(np^2/60) .$$

FSS depends on a pre-specified order of all markers not included in the initial model. This order is either based on some marginal test statistics, or on some conditional score tests as described below. We will thus formally write the fast stepwise search as a function

$$M = \mathrm{FSS}(M_{\mathrm{init}}, \, test, \, criterion) ,$$

to emphasize that it depends on the initial model, on the specific order of markers according to *test*, and on the respective *criterion*. Depending on the order of markers two groups are considered: Group G_1 consists of the best m_1 SNPs, and group G_2 of the best m_2 SNPs. Thus $G_1 \subset G_2$, where G_1 is the set of SNPs along which directed forward steps are performed (see below), whereas SNPs within G_2 might enter the model via so called exchange steps (see below). The exact choice of the parameters m_1 and m_2 turns out to be not too important. The default values of MOSGWA which are also used for simulations are $m_1 = 350$ and $m_2 = 5000$ (as long as $p \geq 5000$).

FSS is based on three algorithmic steps which we call directed forward, exchange, and backward step. FSS starts by considering the initial model M_{init} as the current model. The directed forward step repeatedly tests if enhancing the current model with a SNP decreases the *criterion*, where SNPs within G_1 are considered in the order obtained from the *test* (therefore *directed* forward search). The first SNP which improves the current model is added, and an exchange step follows.

In the exchange step all SNPs in the current model are tested whether exchanging them with suitable other SNPs decreases the *criterion*. Suitable candidates for exchanging SNP S_i are all other SNPs within G_2 whose physical distance to S_i on the chromosome is less than d. The default cutoff value of MOSGWA is $d = 50$, that is we only consider the 49 closest neighboring SNPs in both directions. The algorithm starts with the first SNP of the current model, and tries to exchange it with all suitable candidate SNPs. In case of improvement the first SNP is substituted with the best possible alternative. Then exchanges of the second SNP with all corresponding candidates are performed, and so on till each SNP of the model has been considered once. The idea behind this strategy is that in the directed search step it might happen that not the optimal SNP was chosen, but a correlated SNP might further improve the model. Also after several SNPs have been added to the model it can happen that exchanging a particular SNP of the model is beneficial. Limiting the exchange to SNPs close to S_i which themselves have reasonably large test statistic makes this strategy computationally feasible.

The third step of FSS is an extended backward elimination step. First a standard greedy elimination step is performed, which means that all models are considered where one SNP is removed from the current model. If this does not improve the model, greedy elimination is repeated up to three times to look for better models. In other words like in a classical stepwise elimination procedure we first remove the SNP which explains the least, if this does not improve the selection criterion we additionally remove the next SNP which explains the least, and if the resulting model again has larger selection criterion than the original model we try one more time to remove the least explanatory SNP. The resulting best model is then the starting point for another directed forward step. Directed forward, exchange, and backward steps are then performed repeatedly till no further improvement of the *criterion* is achieved.

Starting with the null model M_0 the complete search strategy of MOSGWA can be specified as follows:

1. $M^* =$ FSS(M_0, Cochran Armitage, mBIC_60)
2. $M^{**} =$ FSS(M^*, Score Test, mBIC_60)
3. $M^{\mathrm{final}} =$ FSS(M^{**}, Score Test, mBIC2)

The general strategy of this algorithm can be motivated as follows. In the first round markers are preselected based on their marginal (Cochran Armitage) test statistic. In the second round score tests conditional on the model M^* of the first round are performed over all remaining SNPs (see [26]). Score tests have the benefit of being computationally much less expensive than likelihood ratio tests, and therefore provide a fast way to preselect markers which might be of importance additional to markers within M^*.

The first two rounds are performed with the milder criterion mBIC_60, which is expected to yield models which are too large. In fact when models are getting too large then the value of d from the exchange step is reduced to guarantee reasonable runtime. As mentioned previously the benefit of first working with mBIC_60 is that one reduces the chances of missing out on important SNPs due to local minima. Only in the final round FSS is performed with the target criterion mBIC2, and will then yield a model for which the type I error rate is controlled in terms of FDR.

The software package MOSGWA is available at http://mosgwa.sourceforge.net/. An application note describing the software in more detail is in preparation.

Simulation Studies

Global null hypothesis

Our first set of simulations is concerned with controlling the type I error under the global null hypothesis. Simulations are based on real SNP data from $n=4077$ individuals from the POPRES sample [34], which are included in the POPRES_Genotypes_QC2 dataset. Individuals are randomly allocated as cases and as controls with equal probability; then MOSGWA, HLASSO and GWASelect are applied to evaluate their ability to control the type I error rate. The random allocation was repeated 200 times, and Table 1 presents the average number of observed false positives to estimate the per-family error rate. A graphical illustration is provided in Figure 1.

To study the influence of the number of SNPs we performed four different simulations using SNPs only from chromosome 1, from the first two, the first four and the first six chromosomes, respectively. The resulting number of SNPs, p, for these four scenarios is given in the first column of Table 1. GWASelect and single marker tests tend to report a large number of correlated SNPs, and thus for these two procedures the number of false positives was obtained by counting clusters of neighboring SNPs [26], where clustering was performed with the algorithm of Frommlet [20]. For HLASSO and MOSGWA no such clustering is necessary, because these algorithms tend to select only one representative for a genomic region anyway.

Table 1 illustrates that MOSGWA controls the type I error under the total null hypothesis irrespective of the number of SNPs. For MOSGWA no parameter needs to be tuned, whereas the latest version of HLASSO allows to choose a parameter α which corresponds to an uncorrected nominal significance level. We consider three Bonferroni-like choices of the form $\alpha \in \{0.3/p, 0.2/p, 0.1/p\}$. The results from Table 1 indicate that the type I error tends to remain below the nominal level, which is not too surprising given the positive correlations between neighboring SNPs due to linkage disequilibrium. Type I error rates for $\alpha=0.3/p$ are closest to those from MOSGWA, and therefore from now on HLASSO will be used with this parameter setting.

The size of the model selected by GWASelect depends on the stable-selection-threshold ξ, which is defined in the following way; GWASelect reports a SNP as detected when the proportion of random subsamples in which the SNP was selected is larger than ξ. He and Lin [26] recommend a choice between 0.1 and 0.2. However, the results from Table 1 show that for these parameter settings GWASelect completely fails to control the type I error rate under the global null. We therefore considered additionally $\xi=0.3$, for which the per-family error rate is controlled at least to some extent. Interestingly the type I error from GWASelect decreases when the number of SNPs increases.

The last column of Table 1 provides the results of single marker tests performed with PLINK [36]. We considered logistic regression models for each marker including the first four principle components of all SNP genotypes as covariates to account for population structure. This kind of adjustment is in principle not necessary when simulating under the total null hypothesis, but it becomes important for the simulation of complex traits below. We applied the Benjamini Hochberg procedure to account for multiple testing, and we can see that under the total null hypothesis the single marker tests control the type I error rate pretty much at the nominal level $\alpha=0.05$.

Complex trait

The second set of simulations is concerned with the power to detect causal SNPs. To this end we consider again the 149478

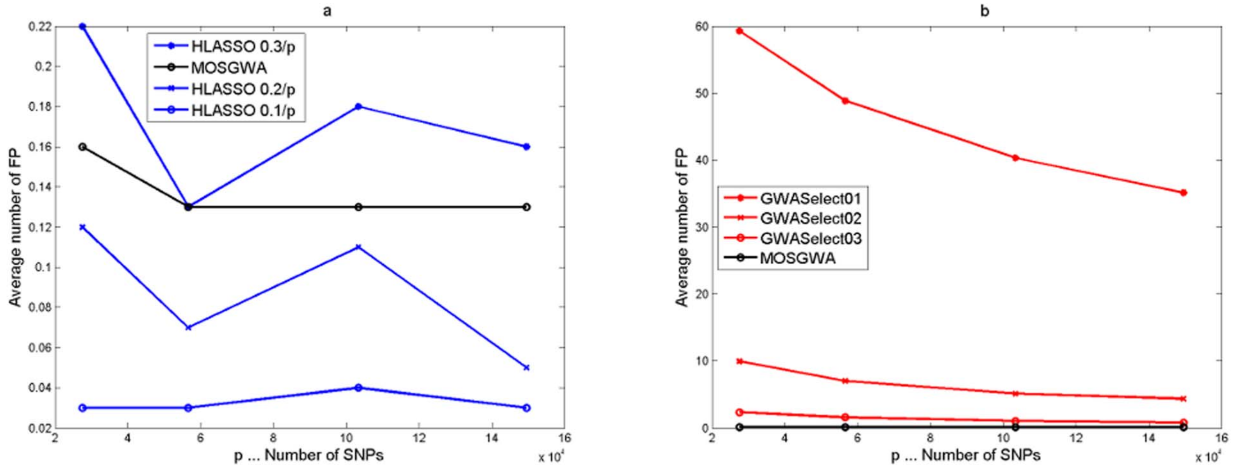

Figure 1. Illustration of the simulation results under the total null hypothesis. The average number of false positives for MOSGWA is compared with HLASSO (Fig 1a) and with GWASelect (Fig 1b), for which false positives were clustered. Simulations were performed for four different numbers of chromosomes, resulting in different numbers of SNPs plotted on the x-axis.
doi:10.1371/journal.pone.0103322.g001

Table 1. False positives under the global null hypothesis.

p	MOS	HLASSO (α)			GWASelect (ξ)			SM
		$.3/p$	$.2/p$	$.1/p$	.1	.2	.3	
27520	0.16	0.22	0.12	0.03	51.1	9.34	2.34	0.04
56629	0.13	0.13	0.07	0.03	43.34	6.73	1.51	0.04
103348	0.13	0.18	0.11	0.04	36.64	4.90	1.06	0.04
149478	0.13	0.16	0.05	0.03	32.22	4.19	0.79	0.06

p refers to the total number of SNPs. The methods analyzed are MOSGWA (MOS), HLASSO with three different choices of the parameter α, GWASelect with three different choices of the stable-selection-threshold ξ, and single marker tests (SM) with Benjamini Hochberg procedure at level $\alpha = 0.05$.

doi:10.1371/journal.pone.0103322.t001

SNPs from the first six chromosomes for the 4077 individuals from the POPRES sample. Simulations are performed for three scenarios, which include 6, 12, and 24 causal SNPs, respectively. All causal SNPs are common (MAF >0.05), equally distributed over the six chromosomes, and with pairwise correlation $\rho < 0.1$ for each pair. Disease risk was computed for each individual according to equation (2), based on which for each scenario 200 case-control data sets were sampled. Effect sizes β_j were ranging in the interval [0.2,0.28], yielding causal SNPs with intermediate power. The coefficient of the intercept β_0 was chosen such that the number of cases and controls in each simulation run was more or less identical.

Before analyzing the data in each scenario half of the causal SNPs were removed, mimicking the situation where SNPs associated with a trait are not causal themselves, but only in linkage disequilibrium with the cause. SNPs to be removed were selected in such a way that there actually were SNPs in linkage disequilibrium, to make it possible for these signals to be detected. The simulated data were then analyzed with MOSGWA, HLASSO (using parameter $\alpha = 0.3/p$), GWASelect using parameters $\xi \in \{0.1, 0.2, 0.3\}$ and with single marker tests (as previously including the four leading principal components in logistic regression models and applying Benjamini Hochberg procedure at nominal level $\alpha = 0.05$).

Table 2 and Figures 2, 3 and 4 summarize the corresponding results in terms of estimated power (which is just the average number of correctly detected signals divided by the total number of causal SNPs), the average number of false positives, the average number of misclassifications (that is false positives plus missed causal SNPs) and the estimated false discovery rate.

A crucial point in computing all these statistics is the definition of true positives and false positives. Of course we know the SNPs which we used to simulate the data, which we will call causal SNPs or correct SNPs. Now there might be several SNPs in close linkage disequilibrium with a causal SNP, and there is the question whether we count a detected SNP which is strongly correlated with the correct SNP as true positive or as false positive. Furthermore half of the causal SNPs under which we simulated were removed before analyzing the data. Therefore to get reasonable results we actually have to count detections which are strongly correlated with a causal SNP as true positives.

It is quite common in this context to use a threshold value C on the correlation between causal SNPs and any detected SNP to determine whether a detected SNP is a true positive [26,29]. If several detected SNPs are closely correlated to one causal SNP we count all of them as one true positive. Just like for the simulation under the total null we have additionally clustered the false positives of GWASelect and of the single marker analysis. This was performed with the algorithm described in Frommlet [20] which computes so called C-clusters, that is clusters of SNPs where it is guaranteed that within each cluster all SNPs have pairwise correlation larger than C. We used for clustering and for determining true positives always the same constant C. Counting the number of clusters rather than the total number of false positives works in favor of the performance of GWASelect and single marker tests. Both for MOSGWA and HLASSO such clustering appears to be unnecessary, because for a genomic region of closely related SNPs usually these procedures select only one representative.

In Figures 2, 3 and 4 we illustrate the dependence of the different statistics on the threshold values C for correlations which specify true positives via $|R| > C$. Results were computed for $C \in \{0.2, 0.3, 0.4, 0.55, 0.7, 0.9\}$. One can see that within the range of $0.2 \leq C \leq 0.55$ the dependence on the threshold is relatively

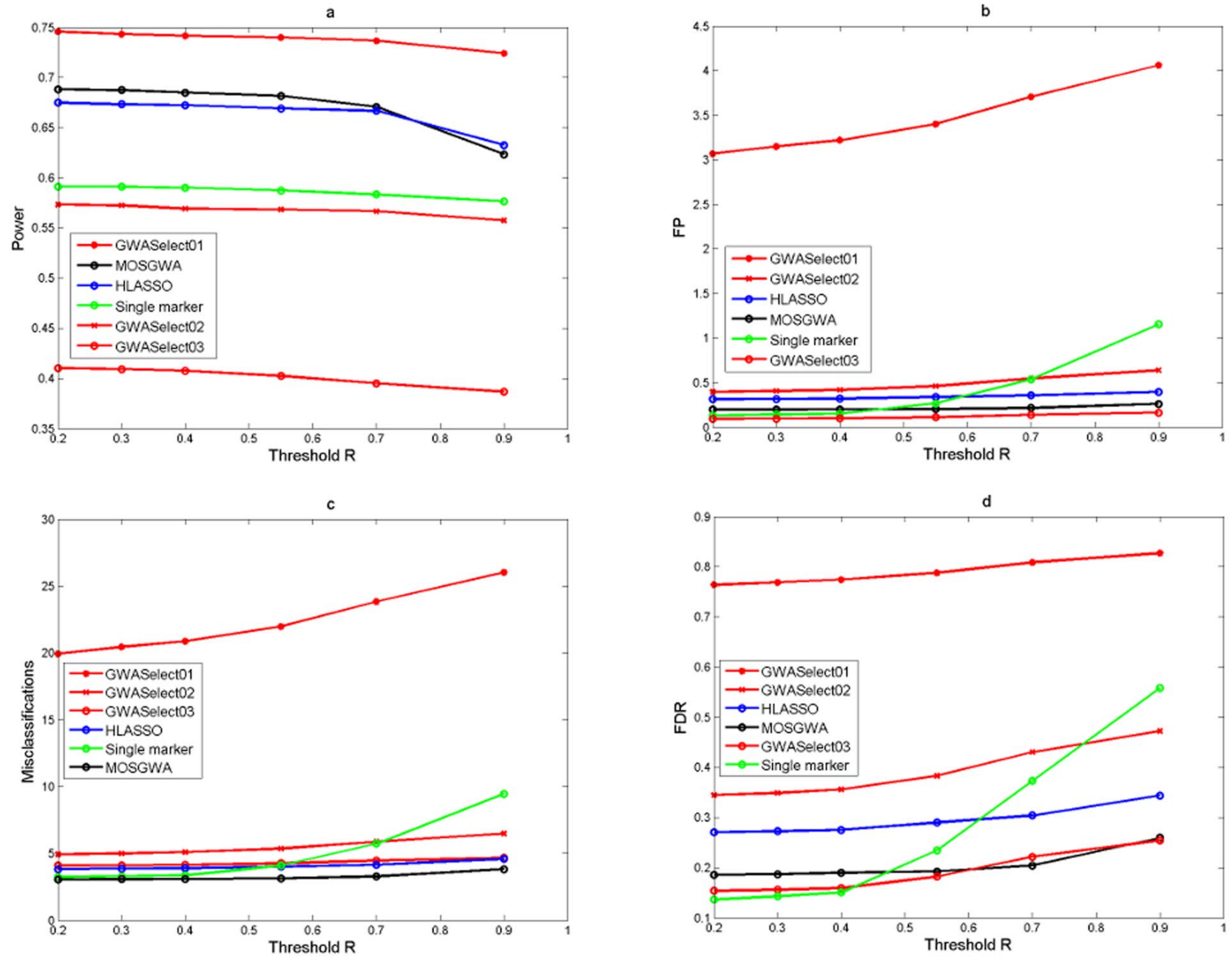

Figure 2. Simulation results under an alternative with $k=6$ causal SNPs. The four panels (Fig 2a, Fig 2b, Fig 2c, Fig 2d) show the average power, number of false positives, misclassification rate and false discovery rate as a function of the threshold value C which determines if a detection is a true or a false positive. The performance of MOSGWA is compared with single marker tests, HLASSO, and with GWASelect using three different parameters for stability selection.
doi:10.1371/journal.pone.0103322.g002

minor for all methods. In general the choice of C has the biggest impact on the results of single marker tests. For those the number of false positives grows much faster than for the other methods, because the number of false positive clusters increases with stricter clustering threshold C. Table 2 specifically reports the results for a threshold $C=0.3$.

For all three scenarios MOSGWA has the lowest number of misclassifications, which is in accordance with theoretical results [21]. Of particular interest is the comparison between MOSGWA and HLASSO. For $k=6$ MOSGWA has slightly larger power and lower Type I error. For $k=12$ and specifically for $k=24$ MOSGWA has much larger power than HLASSO, but also larger Type I error. When the number of causal SNPs is increasing then HLASSO is getting more conservative in comparison with MOSGWA, which is in accordance with the way both algorithms are designed. HLASSO tries to control the FWER at a certain level, whereas MOSGWA tries to control the FDR. According to asymptotic theory [7,24], when aiming at a minimal number of misclassifications in a sparse setting it is preferable to control FDR rather than FWER.

Concerning GWASelect just like under the total null hypothesis the choice of $\xi=0.1$ gives way too large models. The choice of $\xi=0.2$ works slightly better in terms of controlling the Type I error, but is less powerful than MOSGWA and HLASSO. Given the simulation results under the total null hypothesis one actually should use the setting $\xi=0.3$, but then GWASelect is no longer competitive at all in terms of power.

The single marker tests are performing surprisingly well, and the disadvantage compared to the model selection approaches is much less than it was observed by Frommlet et al. [23] for quantitative traits. However, we believe that this is mainly due to the fact that we have added the first four principle components of SNP genotypes. Although we did not specifically simulate scenarios where population structure would play a major role in itself, when testing a specific SNP the principle components in the model help to adjust for the net effect of all the other causal SNPs. Without adding principle components the performance of the single marker tests was extremely poor. Also note that for MOSGWA we did not specifically take any measures to take into account population structure. Still it is much more powerful than single marker tests in all three scenarios, while at the same time controlling FDR at a comparable level.

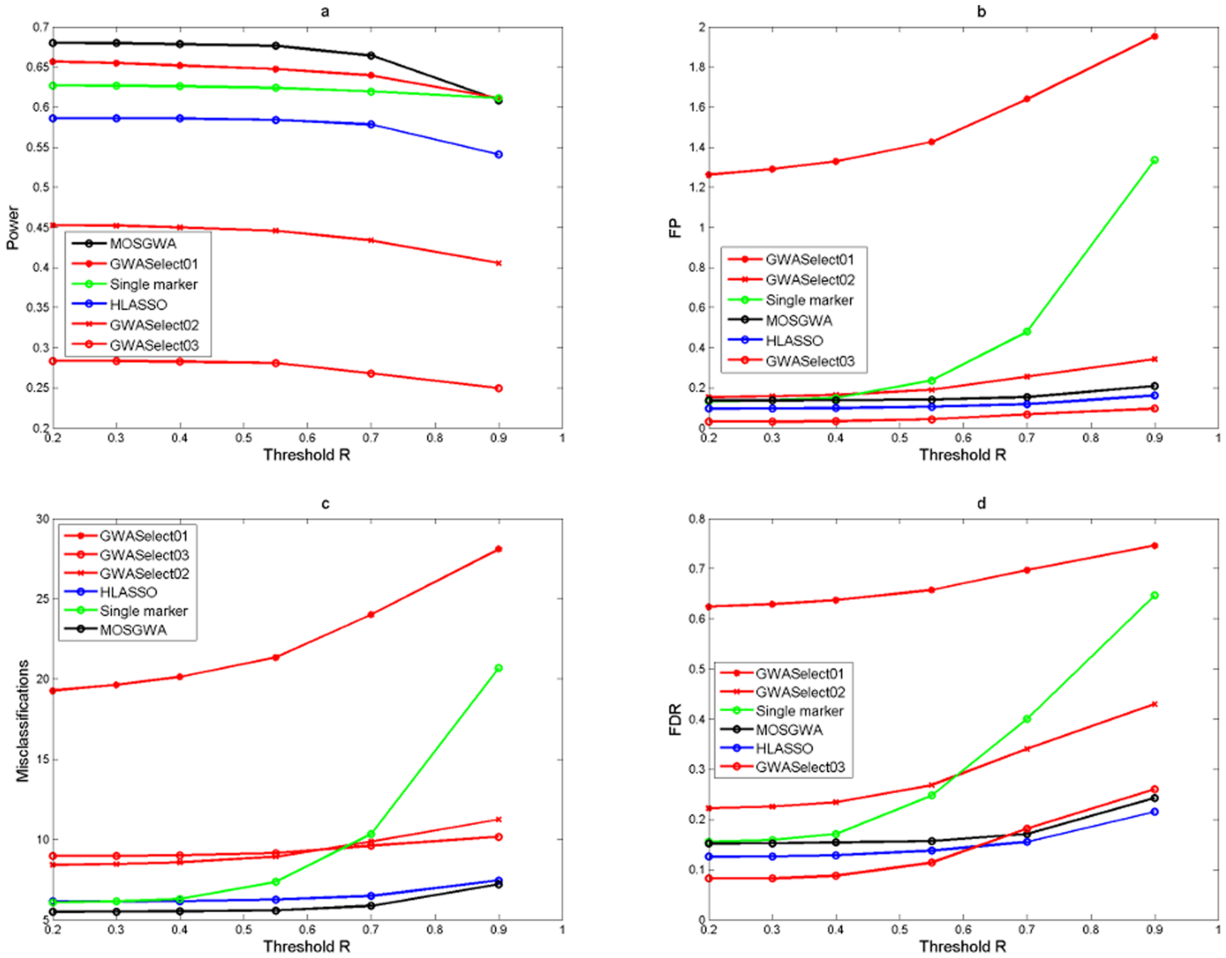

Figure 3. Simulation results under an alternative with $k=12$ causal SNPs. The four panels (Fig 3a, Fig 3b, Fig 3c, Fig 3d) show the average power, number of false positives, misclassification rate and false discovery rate as a function of the threshold value C which determines if a detection is a true or a false positive. The performance of MOSGWA is compared with single marker tests, HLASSO, and with GWASelect using three different parameters for stability selection.
doi:10.1371/journal.pone.0103322.g003

Real Data

The Wellcome Trust data for genome-wide association studies on seven different diseases [41] have become benchmark data sets for comparing different algorithms to analyze GWAS. For each disease approximately 2000 cases were compared with a common set of approximately 3000 controls. More than half a million SNPs were genotyped with the Affymetrix GeneChip 500K, from which less than 400,000 passed quality control. The original analysis from WTCCC was mainly based on single marker tests and identified 24 significant SNPs for the seven diseases.

The analysis of He and Lin [26] using GWASelect resulted in 60 distinct loci. Unfortunately we were not able to completely reproduce their results, which has several reasons. First of all He and Lin [26] did not fully document their preprocessing of data for quality control, and model selection analysis of GWAS data is extremely sensitive with respect to the set of SNPs being studied. Furthermore they seem to have used not always the same parameter ξ for d-GWASelect, but adapted this for different diseases. We provide next a detailed description of the preprocessing steps we performed, and how they differ from [26].

Data preprocessing of WTCCC data

We reanalyzed bipolar disorder (BD), coronary heart disease (CAD), hypertension (HT), Crohn's disease (IBD), rheumatoid athritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). Like in the original article [41] all diseases are compared with the same control group of 3000 individuals. In the following filenames we will denote by <DIS> any of the seven abbreviations for diseases given above. Starting from the already imputed WTCCC data sets we removed individuals for each disease and for the control group according to the files exclusion-list-05-02-2007-<DIS>.txt and exclusion-list-snps-26_04_2007.txt.

After merging cases with controls we removed SNPs with a minimal allelic frequency smaller than 0.01. Furthermore we tested for Hardy Weinberg equilibrium and SNPs with p-values smaller than 0.0001 were also removed. Finally we only considered SNPs for which the genotype calling algorithm from the WTCCC confirmed good clustering, where we took that information from the files WTCCC_summary_data/7_Diseases/<DIS>/basic/snptest_<DIS>_<CHR>.txt. Here <CHR> is the chromosome number ranging from 1 to 22. According to private correspondence with He and Lin they had used the file wtccc_<DIS>_basic_chr_<CHR>.xml for similar purposes. That file must have differed from the file we have used, but is no

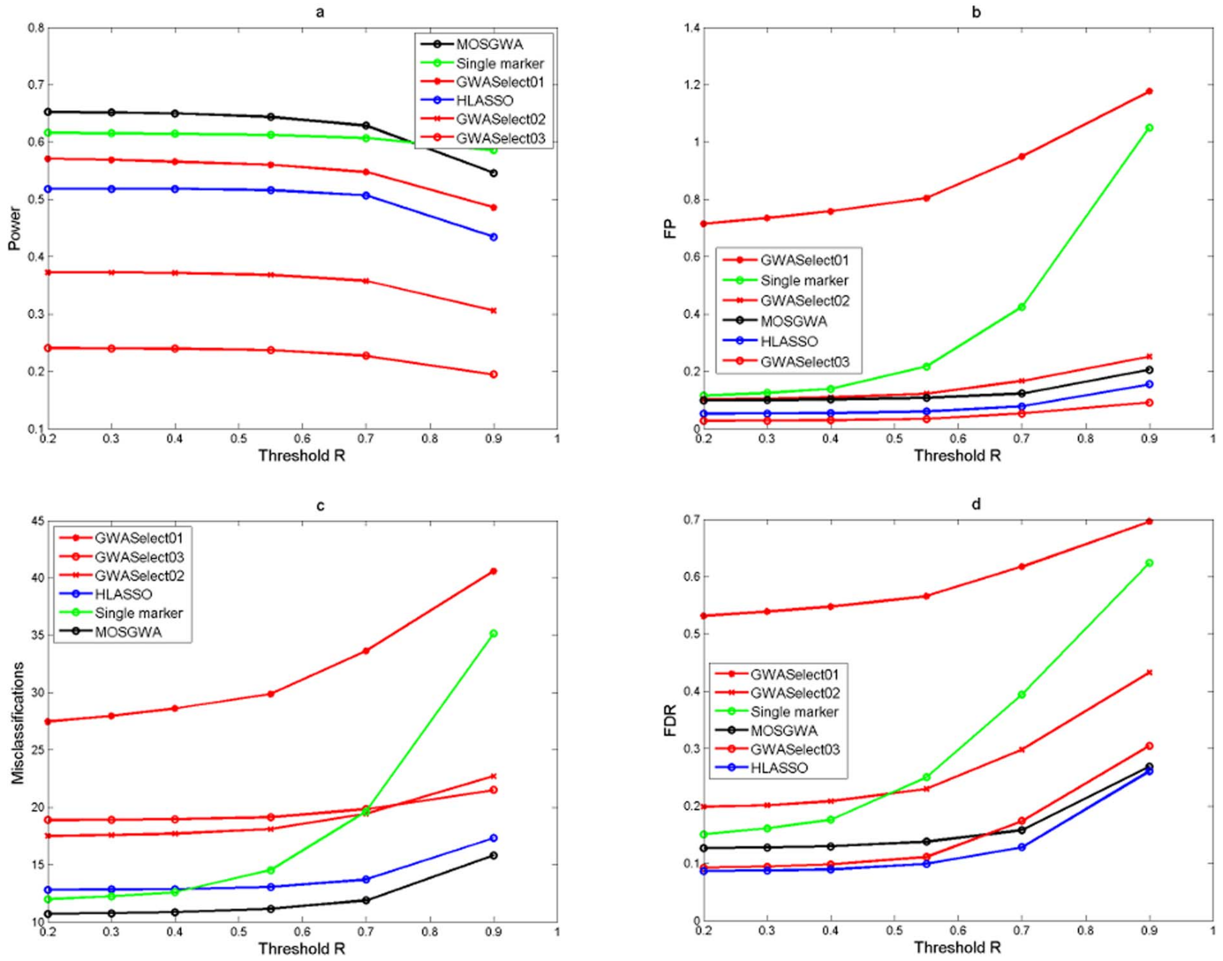

Figure 4. Simulation results under an alternative with $k=24$ causal SNPs. The four panels (Fig 4a, Fig 4b, Fig 4c, Fig 4d) show the average power, number of false positives, misclassification rate and false discovery rate as a function of the threshold value C which determines if a detection is a true or a false positive. The performance of MOSGWA is compared with single marker tests, HLASSO, and with GWASelect using three different parameters for stability selection.
doi:10.1371/journal.pone.0103322.g004

longer available from WTCCC. In summary, after preprocessing we will have ended up with a different set of SNPs than used in [26], and according to email correspondence with the authors it is no longer possible to reconstruct which set of SNPs they had used for their own analysis. Not starting from the same set of SNPs explains why we were not able to completely reproduce the results reported by He and Lin [26].

Summary of results

Like in the simulation study for complex traits HLASSO was applied with parameter $\alpha=0.3$. For GWASelect we present the results again for $\xi \in \{0.1, 0.2, 0.3\}$, in spite of the fact that we have seen that only $\xi=0.3$ controls the type I error rate under the total null hypothesis. Table 3 gives for each disease the number of detected SNPs (and associated regions) obtained from the original WTCCC analysis, MOSGWA, HLASSO, and GWASelect. A more detailed discussion concerning the detected SNPs follows below.

The first observation is that GWASelect with parameter $\xi=0.1$ is selecting in 6 out of 7 diseases a much larger number of SNPs than the other methods. Given the results from simulations we are forced to conclude that most of those SNPs might be false positives, and we will not give the detailed results for GS 0.1 in the next section except for rheumatoid arthritis, where interestingly GS 0.1 gives quite similar results to MOSGWA and HLASSO (for a detailed explanation see below). For Crohn's disease and Type I diabetes the results for $\xi=0.2$ are relatively close to MOSGWA and HLASSO, whereas for all other diseases $\xi=0.3$ might be the best choice for GWASelect. The general conclusion is that the results of GWASelect heavily depend on the choice of ξ, and it is not really possible to know in advance which choice gives reliable results.

The comparison between MOSGWA and HLASSO is quite interesting. For the four diseases for which only a small number of SNPs was detected (BD, CAD, HT, T2D) MOSGWA finds exactly the same regions as HLASSO, though in two cases one SNP less. Similarly MOSGWA has a tendency to select less representatives of a region than HLASSO for the remaining more complex traits. This might have to do with the fact that the coefficient estimates of MOSGWA suffer from even less shrinkage than the estimates from HLASSO. Hoggart et al. [29] provide a thorough discussion of the fact that in case of shrinkage the regressors which enter a model explain less than they would without shrinkage, which results in a higher chance of including further correlated SNPs in the model. This is the main reason why HLASSO works with the NEG prior, which results in less

Table 2. Summary of simulation results for complex traits.

	MOS	HL	GS 0.1	GS 0.2	GS 0.3	SM
Scenario 1: ($k = 6$)						
Size	5.32	5.94	23.36	5.87	3.03	4.40
Power	0.69	0.67	0.74	0.57	0.41	0.59
FP	1.19	1.90	18.90	2.43	0.57	0.85
FDR	0.19	0.27	0.77	0.35	0.16	0.14
Mis	3.07	3.86	20.44	5.00	4.12	3.31
Scenario 2: ($k = 12$)						
Size	9.80	8.19	23.35	7.31	3.77	9.17
Power	0.68	0.59	0.65	0.45	0.28	0.63
FP	1.65	1.16	15.49	1.88	0.37	1.65
FDR	0.15	0.17	0.63	0.23	0.08	0.16
Mis	5.49	6.13	19.63	8.46	8.97	6.13
Scenario 3: ($k = 24$)						
Size	18.03	13.71	31.29	11.43	6.43	17.76
Power	0.65	0.52	0.57	0.37	0.24	0.62
FP	2.39	1.27	17.63	2.49	0.67	2.99
FDR	0.13	0.09	0.54	0.20	0.09	0.16
Mis	10.75	12.83	27.97	17.56	18.91	12.22

The average over 200 simulation runs is reported for the number of detected associations (Size), the estimated power, the number of false positive detections (FP), the estimated false discovery rate (FDR) and the average number of misclassifications (Mis). GWASelect performed with parameters $\xi \in \{0.1, 0.2, 0.3\}$ is abbreviated as GS ξ, MOSGWA as MOS, HLASSO as HL, and single marker tests as SM.

doi:10.1371/journal.pone.0103322.t002

shrinkage than the double exponential prior corresponding to LASSO.

A second observation is that especially for complex traits MOSGWA tends to select more regions of association than HLASSO. This goes along with the fact that HLASSO is designed to control the FWER, whereas MOSGWA controls the FDR. In case of complex traits MOSGWA is therefore bound to find more SNPs than HLASSO, whereas if there are only few signals both methods behave very similarly. Looking more closely into the results for IBD, RA and T1D, we observe that MOSGWA detects SNPs within all regions which were reported as significant according to the original WTCCC analysis [41]. The same is not true for HLASSO, which misses out one region on chromosome 10 for RA, and another region on chromosome 16 for T1D. On the other hand MOSGWA finds exclusively 7 SNPs for IBD, 4 SNPs for RA, and 6 SNPs for T1D, respectively. We will next provide a thorough discussion on the potential relevance of these SNPs which were exclusively found by MOSGWA.

WTCCC Results in detail

For each disease we provide tables as supporting information which list all SNPs which were detected by different methods. Specifically the first column of these tables gives the reference SNP ID number from dbSNP, followed by the chromosome (Chr) and the position (Pos). The column Gene contains information about the closest lying gene according to the databases dbSNP and ImmunoBase:

Table 3. Summary of real data analysis.

Disease	WTCCC	MOS	HL	GS 0.3	GS 0.2	GS 0.1
BD	(1)	1 (1)	1 (1)	1 (1)	9 (9)	43 (39)
CAD	(1)	2 (2)	3 (2)	3 (2)	4 (2)	29 (21)
HT	(0)	1 (1)	1 (1)	1 (1)	4 (4)	29 (26)
IBD	(9)	17 (16)	12 (8)	12 (5)	15 (6)	32 (19)
RA	(3)	11 (5)	12(2)	1(1)	1(1)	13 (2)
T1D	(7)	25 (11)	22 (4)	12 (2)	20 (2)	33 (2)
T2D	(3)	2 (2)	3 (2)	4 (2)	8 (4)	28 (19)

Number of detected SNPs which are associated to the following seven diseases from WTCCC: Bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), Crohn's disease (IBD), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). WTCCC refers to the regions reported by the original publication [41] in their Table 3, abbreviations for the other algorithms are just like in Table 2. In brackets we give the number of DNA regions which are covered by the detected SNPs. The whole HLA region on chromosome 6 is counted as only one region.

doi:10.1371/journal.pone.0103322.t003

- dbSNP: http://www.ncbi.nlm.nih.gov/SNP/
- ImmunoBase: http://www.immunobase.org/

The next four columns have bullets whenever a SNP was detected by MOSGWA (M), Hlasso (HL), GWASelect with parameter $\xi = 0.3$ (G3) or $\xi = 0.2$ (G2). According to our simulation results we believe that the large number of additional SNPs detected by GWASelect with $\xi = 0.1$ (G1) will include mainly false positives. Therefore we do not present detailed results for G1, with the exception of rheumatoid arthritis which has a rather particular genetic constellation. If many neighboring SNPs are reported by different methods we consider such groups of SNPs as genetic regions, and we label such groups in the tables using background colors. The crosses in the last column (W) indicate regions which were reported originally in [41].

Bipolar disorder (BD). As illustrated in table S1 SNP rs2837588, which is an intron of DSCAM, was found by all methods, though it was not reported in [41]. On the other hand the only SNP reported by WTCCC, rs4202459, was not detected by any of the algorithms we analyzed here. This can be quite easily explained because all algorithms we study here are based on models incorporating linear trends, and the trend p-value of rs4202459 is quite large (2.19 E-04 according to the WTCCC manuscript). That rs2837588 was not reported by WTCCC might have to do with the large number of missing values for this SNP. Results reported in [41] are not based on imputed data, and imputation changes the marginal p-value for this SNP considerably. More recent research [3] indicates that there actually might be a connection between bipolar disorder and DSCAM, although in general according to a recent large GWAS [37] it appears to be extremely difficult to identify robust and replicable genetic causes for psychiatric disorders. Thus all the other SNPs reported by G2 have a good chance to be false positives.

Coronary artery disease (CAD). Here the original WTCCC study reported only one region, whereas all methods studied here detect two regions (see table S2). Note that MOSGWA selects only one representative of the region reported by WTCCC, whereas HLASSO reports two representatives. This is a pattern we will see again several times, for example in type II diabetes. A possible explanation why HLASSO prefers to choose more representatives of a region than MOSGWA is that MOSGWA is based on model selection criteria which impose less shrinkage on the coefficients than HLASSO does. The effect of shrinkage on the number of selected correlated SNPs is thoroughly described by Hoggart et al. [29].

Hypertension (HT). The original WTCCC study did not report any SNP associated with hypertension, whereas all methods studied here report rs16945811 on chromosome 17 (table S3). The reason for this is again that we work with imputed data, for which the marginal p-value of this SNP is considerably smaller than for the original unimputed data. The other SNPs reported by G2 are again very likely to be false positives.

Crohn's disease (IBD). This is the first disease for which model selection approaches become really interesting, because the trait appears to be a complex one. Note that HLASSO finds one region less (on chromosome 3) than originally reported in [41], whereas MOSGWA finds all regions reported by WTCCC plus seven additional ones which are highlighted in yellow in the table S4. At least five of those have been mentioned meanwhile in the literature on Crohn's disease, which means that based on the WTCCC data MOSGWA would have detected a number of SNPs associated with Crohn's disease which were later confirmed by independent studies.

Let's look at those SNPs in detail. The first SNP rs12035082 is close to rs12037606 which was actually reported in [41] among the SNPs which showed moderate evidence of association, and was later confirmed to be associated with CD [32]. Similarly rs6908425 was reported as being moderately associated in [41], and could later be confirmed in an independent study [4]. rs4263839 was not reported by WTCCC, but it was among the list of confirmed SNPs given in [4]. Furthermore it has later been shown to be associated with irritable bowel syndrome [44]. rs2836753 is in close linkage disequilibrium with rs2836754, which was related to Crohn's disease in [35].

rs6908425 is an intron from the CDKALI gene on chromosome 6 and has been confirmed to be associated to Crohn's disease (see for example snpedia at http://snpedia.com/index.php/Rs6908425). rs9405639 on chromosome 9 lies within the intron of SLC22A23 gene, which is also well known to be related to Crohn's disease (http://www.immunobase.org/page/Overview/display/gene_id/63027).

The only two SNPs detected by MOSGWA based on the WTCCC data which have not been confirmed in the literature are rs11627513 and rs41526044. rs11627513 is relatively close to the IL23R gene which is known to be associated with Crohn's disease (see http://www.immunobase.org/page/Overview/display/gene_id/149233) and it was mentioned in [28] as a potentially associated gene. Up to our knowledge only for rs11627513 nothing is known, and this might well be a false positive. Remember that MOSGWA is designed to control the FDR approximately at a level of 10%, and thus one would actually expect 2 false positive SNPs within this model.

Rheumatoid arthritis (RA). All SNPs detected on chromosome 6 belong to the so called HLA region which has been well known for a long time to be associated with rheumatoid arthritis (see for example [15]). However, HLA genes explain only approximately one-third of the genetic liability of the disease [13], and a great amount of research has been performed to understand genetic causes beyond HLA. For the rheumatoid arthritis data GWASelect performs very poor, for parameters $\xi = 0.3$ and $\xi = 0.2$ only one SNP in the HLA region is reported. This is quite easy to understand given the way GWASelect works. In the HLA region there are many highly correlated SNPs, and during the stability selection procedure it is very likely that for different samples different representatives of a cluster of SNPs are chosen. Therefore only for the lowest threshold $\xi = 0.1$ GWASelect gives results which are more in line with the other methods. This is the reason why He and Lin [26] clustered the HLA region before analyzing that data set, but that appears to be quite an extra effort for the user when applying GWASelect. Neither MOSGWA nor HLASSO did have particular problems with including the HLA region in the analysis, and both find apart from SNPs in the HLA region also rs6679677 on chromosome 1. This SNP was already reported by the WTCCC [41], but according to a recent meta-analysis [13] this region actually could not be confirmed to be associated with rheumatoid arthritis.

Four SNPs were then only detected by MOSGWA which are highlighted in the table S5 in yellow. These detections appear to be rather interesting. rs2104286 lies in the intron of the IL-2RA gene, and according to a recent study [13] this SNP is definitely associated with rheumatoid arthritis. It was reported by WTCCC [41] after pooling data from RA and T1D, but it was not detected when using the standard analysis for data only from the RA population. rs1946518 from chromosome 11 is not mentioned in the meta-analysis, but the GUCY1A2 gene is down regulated in case of rheumatoid arthritis [16]. This indicates that there might be a functional connection between this region and RA.

Finally there remain two SNPs from the GLI3 gene on chromosome 7. Again nothing is said about those in the meta-analysis [13]. However, the GLI3 gene is known to be a member of the Hedgehog signaling pathway which is important for the proper development of embryos, and is also known to play an important role in adults [40]. rs12536071 has a relatively small marginal p-value (5.2E-06), whereas the neighboring SNP rs12531052 has a rather large marginal p-values (0.167944). It is pretty unusual for MOSGWA to select two SNPs which are located so close to each other, which might indicate that actually some epistatic effect could be involved here.

Type I diabetes (T1D). Similarly to rheumatoid arthritis also for type I diabetes the HLA region plays an important role, where approximately half of the genetic risk for T1D is found in the HLA region [10]. GWASelect is handling the situation here slightly better than in case of RA, but still it detects apart from HLA SNPs only rs6679677 on chromosome 1, which is also reported by all other methods. In the original WTCCC publication [41] three more SNPs are found to have strong association with T1D. All of those were found by MOSGWA, whereas HLASSO missed rs12924729 on chromosome 16. MOSGWA reported 6 additional SNPs which are again highlighted in yellow in the table S6.

rs41384747 on chromosome 18 is close to rs2542151 which was already listed among SNPs with moderate association by WTCCC [41]. The region was later confirmed in replication studies (see [5]), and also includes SNP rs478582 listed in the meta-analysis [10]. Similarly the region on chromosome 5 in which rs1025039 lies has been known as a susceptibility locus for T1D. In the same region lies rs6897932 which is associated with T1D [5]. Perhaps the most comprehensive source today for human Type 1 Diabetes loci is the database http://t1dbase.org. There one can find that the region on chromosome 18 is listed as being associated with T1D, but not the region on chromosome 5.

The closest region to rs6928921 on chromosome 6 which is documented in t1dbase is 6q15 with the BACH2 gene, which is 5 MB upstream. Nothing is known about any influence of KIAA1009 on T1D. There is more indication that rs2666236 on chromosome 10 might be associated with T1D, as it is known that the corresponding gene NRP1 is associated with T1D [25]. However, again the region of rs2666236 is not listed in t1dbase. Nothing is known about the other SNPs rs7157296 on chromosome 14 and rs41384747 on chromosome 18, and they might thus be false positives. Remember that MOSGWA is tuned to have an FDR of approximately 10%, and thus 2 or 3 false positive SNPs are to be expected in this model.

Type II diabetes (T2D). With type 2 diabetes a smaller number of genetic regions seems to be associated than with the previous three diseases (table S7). Thus the model selection approach appears to have less benefits compared with the standard analysis. MOSGWA does not report rs9465871 on chromosome 6, but looking at the original WTCCC analysis [41] this SNP has a trend p-value of 1.02 E-6, which is not significant after any standard correction for multiple testing. Apart from that all algorithms report SNPs within the two main regions on chromosome 10 and 16, where MOSGWA chooses again only one representative for gene TCF7L2 on chromosome 10.

Discussion

We have introduced MOSGWA, a new algorithm for GWAS analysis using the FDR controlling model selection criterion mBIC2. We compared its performance with two existing variable selection methods, GWASelect and HLASSO. The first observation was that both MOSGWA and HLASSO are controlling the Type I error rate under the global null hypothesis, whereas GWASelect does not manage to do that when using the recommended parameter setting $\xi \in [0.1, 0.2]$. He and Lin [26] only presented simulation results for scenarios including 10 causal SNPs, for which GWASelect had relatively low FDR. In accordance we observed that GWASelect tends to have lower type I error rate when the true model underlying simulations includes more causal SNPs. However, a method which selects very large models even when we know that there is no genetic cause for the disease status appears to be rather problematic. When increasing the selection threshold to $\xi = 0.3$ then GWASelect more or less controls the type I error under the global null, but then it is no longer competitive in terms of power to detect causal SNPs. One advantage of MOSGWA is that there is actually no parameter tuning necessary at all, because the selection criterion is fixed.

Apart from GWASelect and HLASSO we have also compared the performance of the original LASSO as implemented in glmnet [19], although we decided not to present the corresponding results in this manuscript. LASSO previously has been shown to perform well in GWAS in terms of prediction [30], but here we are only interested in selecting the correct SNPs. For this purpose one has the problem to decide upon the best tuning parameter λ of LASSO. It is well known that cross validation yields too large models, which we also observed in our simulations. As an alternative we tried to search along the LASSO regularization path and find the model which minimizes mBIC2, but that gave too small models. We finally considered a strategy where we searched along the regularization path for that model which minimizes the misclassification rate. This is obviously not feasible in practice when the truth is unknown, but this strategy shows the best possible performance of LASSO that one could achieve at least theoretically. It still turned out that in our simulations even the best possible model along the regularization path could not compete with the models obtained with MOSGWA or HLASSO. We observed that LASSO tends to select too many correlated SNPs with larger effect sizes, and then has difficulties to include causal SNPs with smaller effect size, but rather includes a number of false positives. An explanation of this behavior was given already by Hoggart et al. [29] when motivating the NEG prior.

Our simulation study on complex traits showed that at least for our three scenarios MOSGWA is slightly more powerful than HLASSO, when the parameter α of HLASSO is chosen such that both procedures have similar type 1 error rate. This reflects the theoretical optimality property of mBIC2 to minimize asymptotically the misclassification error under a wide range of sparsity levels [21]. However, this theoretical property holds for the model which actually minimizes the criterion mBIC2. Our heuristic search strategy is attempting to get close to the global minimum, but we know that in most cases it will fail to find the best solution. More involved search strategies will further improve our method, and we are currently exploring the use of memetic algorithms, which have been successfully applied already in the context of QTL mapping [22].

The software architecture of MOSGWA is designed in such a way that it can be quite easily extended in the future to incorporate more advanced features. For example it might be interesting to have a model selection procedure which accounts for population structure. Currently this can be done by adding principle components as covariates to the regression models, but an even better solution would be to add random effects to model population structure. For that reason we are currently working on extending MOSGWA towards mixed models.

Supporting Information

Table S1 Summary of analysis results for Bipolar Disorder.
(PDF)

Table S2 Summary of analysis results for Coronary artery disease.
(PDF)

Table S3 Summary of analysis results for Hypertension.
(PDF)

Table S4 Summary of analysis results for Crohn's disease.
(PDF)

Table S5 Summary of analysis results for Rheumatoid arthritis.
(PDF)

Table S6 Summary of analysis results for Type I diabetes.
(PDF)

Table S7 Summary of analysis results for Type II diabetes.
(PDF)

Acknowledgments

The data set used for simulations in this manuscript was obtained from dbGaP through dbGaP accession number phs000145.v1.p1 at www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id = phs000145.v1.p1.

The WTCCC data is available at http://www.wtccc.org.uk/.

Author Contributions

Conceived and designed the experiments: ED BB FF. Performed the experiments: ED. Analyzed the data: ED. Contributed reagents/materials/analysis tools: ED BB FF. Contributed to the writing of the manuscript: ED BB FF.

References

1. Abramovich F, Benjamini Y, Donoho DL, Johnstone IM (2006) Adapting to unknown sparsity by controlling the false discovery rate. Ann Stat 34: 584–653.
2. Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71: 1–10.
3. Amano K, Yamada K, Iwayama Y, Detera-Wadleigh SD, Hattori E, et al. (2008) Association study between the Down syndrome cell adhesion molecule (DSCAM) gene and bipolar disorder. Psychiatr Genet 18(1): 1–10. doi: 10.1097/YPG.0b013e3281ac238e.
4. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40(8): 955–962. doi: 10.1038/ng.175.
5. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41(6): 703–707. doi: 10.1038/ng.381.
6. Benjamini Y, 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.
7. Bogdan M, Chakrabati A, Frommlet F, Ghosh JK (2011) Asymptotic Bayes-Optimality under sparsity of some multiple testing procedures. Ann Stat 39(3): 1551–1579.
8. Bogdan M, Ghosh JK, Doerge RW (2004) Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitive trait loci. Genetics 167: 989–999.
9. Bogdan M, Żak-Szatkowska M, Ghosh JK (2008) Selecting explanatory variables with the modified version of Bayesian Information Criterion. Qual Reliab Eng Int 24: 627–641.
10. Bradfield JP, Qu HQ, Wang K, Zhang H, Sleiman PM, et al. (2011) A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. Plos Genet 7(9): e1002293. doi: 10.1371/journal.pgen.1002293.
11. Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J Roy Statist Soc Ser B 64(4): 641–656.
12. Bühlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer (Heidelberg, London, New York).
13. Chatzikyriakidou A, Voulgari PV, Lambropoulos A, Drosos AA (2013) Genetics in rheumatoid arthritis beyond HLA genes: what meta-analyses have shown? Semin Arthritis Rheu 43(1): 29–38. doi: 10.1016/j.semarthrit.2012.12.003.
14. Chen J, Chen Z (2008) Extended Bayesian Information criteria for model selection with large model spaces. Biometrika 95(3): 759–771.
15. Deighton CM, Walker DJ, Griffiths ID, Roberts DF (1989) The contribution of HLA to rheumatoid arthritis. Clin Genet 36: 178–182.
16. Del Rey MJ, Izquierdo E, Usategui A, Gonzalo E, Blanco FJ, et al. (2010) The transcriptional response of normal and rheumatoid arthritis synovial fibroblasts to hypoxia. Arthritis Rheum 62(12): 3584–3594. doi: 10.1002/art.27750.
17. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J Roy Statist Soc Ser B 70: 849–911.
18. Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80: 27–38.
19. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33(1): 1–22.
20. Frommlet F (2010) Tag SNP selection based on clustering according to dominant sets found using replicator dynamics. Adv Data Anal Classif 4: 65–83.
21. Frommlet F, Chakrabarti A, Murawska M, Bogdan M (2011) Asymptotic Bayes optimality under sparsity for generally distributed effect sizes under the alternative. Technical report: arXiv:1005.4753.
22. Frommlet F, Ljubic I, Arnardottir HB, Bogdan M (2012) QTL mapping using a memetic algorithm with modifications of BIC as fitness function. Stat Appl Genet Mol Biol 11(4): Article 2. doi: 10.1515/1544-6115.1793.
23. Frommlet F, Ruhaltinger F, Twarog P, Bogdan M (2012) Modified versions of Bayesian Information Criterion for genome-wide association studies. Comput Stat Data An 56(5): 1038–1051.
24. Frommlet F, Bogdan M (2013) Some optimality properties of FDR controlling rules under sparsity. Electronic Journal of Statistics 7: 1328–1368, doi: 10.1214/13-EJS808.
25. Hasan NM, Kendrick MA, Druckenbrod NR, Huelsmeyer MK, Warner TF, et al. (2010) Genetic association of the neuropilin-1 gene with type 1 diabetes in children: Neuropilin-1 expression in pancreatic islets. Diabetes Res Clin Pract 87(3): e29–32. doi: 10.1016/j.diabres.2009.12.016.
26. He Q, Lin D (2011) A variable selection method for genome-wide association studies. Bioinformatics 27(1): 1–8.
27. Heinze G, Schemper M (2002) A Solution to the Problem of Separation in logistic regression. Stat Med 21: 2409–2419.
28. Hoffman GE, Logsdon BA, Mezey JG (2013) PUMA: a unified framework for penalized multiple regression analysis of GWAS data. Plos Comput Biol 9(6): e1003101. doi: 10.1371/journal.pcbi.1003101.
29. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (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.
30. Kooperberg C, LeBlanc M, Obenchain V (2010) Risk Prediction Using Genome-Wide Association Studies. Genet Epidem 34: 643–652.
31. Li J, Das K, Fu G, Li R, Wu R (2010) The Bayesian Lasso for Genome-wide Association Studies. Bioinformatics 27(4): 516–523.
32. Weersma RK, Stokkers PC, Cleynen I, Wolfkamp SC, Henckaerts L, et al. (2009) Confirmation of multiple Crohn's disease susceptibility loci in a large Dutch-Belgian cohort. Am J Gastroenterol 104(3): 630–638. doi: 10.1038/ajg.2008.112.
33. Meinshausen N, Bühlmann P (2010) Stability Selection (with discussion) J Roy Statist Soc Ser B 72: 417–473.
34. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, et al. (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet: 83(3): 347–358.
35. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, et al. (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39(7): 830–832. Epub 2007 Jun 6.
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81: 559–575.
37. Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, et al. (2013) A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatr 18: 497–511. doi:10.1038/mp.2012.21.
38. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6, 461–464.
39. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Statist Soc Ser B 58(1): 267–288.
40. Villavicencio EH, Walterhouse DO, Iannaccone PM (2000) The sonic hedgehog-patched-gli pathway in human development and disease. Am J Hum Genet 67(5): 1047–1054.
41. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.

42. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2011) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25(6): 714–721.
43. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7): 565–608.
44. Zucchelli M, Camilleri M, Andreasson AN, Bresso F, Dlugosz A, et al. (2011) Association of TNFSF15 polymorphism with irritable bowel syndrome. Gut 60(12): 1671–1677. doi: 10.1136/gut.2011.241877.
45. Żak-Szatkowska M, Bogdan M (2011) Modified versions of Bayesian Information Criterion for sparse Generalized Linear Models. Comput Stat Data An 55: 2908–2924.