Assessing statistical significance in multivariable genome wide association analysis

Buzdugan, Laura ; Kalisch, Markus ; Navarro, Arcadi ; Schunk, Daniel ; Fehr, Ernst ; Bühlmann, Peter

Abstract: Motivation: Although Genome Wide Association Studies (GWAS) genotype a very large number of single nucleotide polymorphisms (SNPs), the data is often analyzed one SNP at a time. The low predictive power of single SNPs, coupled with the high significance threshold needed to correct for multiple testing, greatly decreases the power of GWAS. Results: We propose a procedure in which all the SNPs are analyzed in a multiple generalized linear model, and we show its use for extremely high-dimensional datasets. Our method yields p-values for assessing significance of single SNPs or groups of SNPs while controlling for all other SNPs and the family wise error rate (FWER). Thus, our method tests whether or not a SNP carries any additional information about the phenotype beyond that available by all the other SNPs. This rules out spurious correlations between phenotypes and SNPs that can arise from marginal methods because the "spuriously correlated" SNP merely happens to be correlated with the "truly causal" SNP. In addition, the method offers a data driven approach to identifying and refining groups of SNPs that jointly contain informative signals about the phenotype. We demonstrate the value of our method by applying it to the seven diseases analyzed by the WTCCC (The Wellcome Trust Case Control Consortium, 2007). We show, in particular, that our method is also capable of finding significant SNPs that were not identified in the original WTCCC study, but were replicated in other independent studies.

DOI: https://doi.org/10.1093/bioinformatics/btw128

BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–11

Assessing statistical significance in multivariable genome wide association analysis

Laura Buzdugan [1,2], Markus Kalisch [1], Arcadi Navarro [3,4,5], Daniel Schunk [6], Ernst Fehr [2] and Peter Bühlmann [1*]

[1]Seminar for Statistics, Department of Mathematics, ETH Zürich, Rämistrasse 101, 8092, Zürich, Switzerland [2]Department of Economics, University of Zürich, Blümlisalpstrasse 10, 8006, Zürich, Switzerland [3]Institute of Evolutionary Biology (CSIC-UPF), Universitat Pompeu Fabra, 08003, Barcelona, Spain [4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona Biomedical Research Park (PRBB), 08003, Barcelona, Spain [5]Center for Genomic Regulation (CRG), Barcelona Biomedical Research Park (PRBB), 08003, Barcelona, Spain [6]Department of Economics, University of Mainz, Jakob-Welder-Weg 4, Mainz, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Although Genome Wide Association Studies (GWAS) genotype a very large number of single nucleotide polymorphisms (SNPs), the data is often analyzed one SNP at a time. The low predictive power of single SNPs, coupled with the high significance threshold needed to correct for multiple testing, greatly decreases the power of GWAS.

Results: We propose a procedure in which all the SNPs are analyzed in a multiple generalized linear model, and we show its use for extremely high-dimensional datasets. Our method yields p-values for assessing significance of single SNPs or groups of SNPs while controlling for all other SNPs and the family wise error rate (FWER). Thus, our method tests whether or not a SNP carries any additional information about the phenotype beyond that available by all the other SNPs. This rules out spurious correlations between phenotypes and SNPs that can arise from marginal methods because the "spuriously correlated" SNP merely happens to be correlated with the "truly causal" SNP. In addition, the method offers a data driven approach to identifying and refining groups of SNPs that jointly contain informative signals about the phenotype. We demonstrate the value of our method by applying it to the seven diseases analyzed by the WTCCC (The Wellcome Trust Case Control Consortium, 2007). We show, in particular, that our method is also capable of finding significant SNPs that were not identified in the original WTCCC study, but were replicated in other independent studies.

Availability: Reproducibility of our research is supported by the open-source Bioconductor package hierGWAS.

Contact: peter.buehlmann@stat.math.ethz.ch

1 INTRODUCTION

Genome-Wide Association Studies (GWAS) have enjoyed increasing success and popularity in recent years, due mostly to the thousands of genetic variants found to be significantly associated with complex traits (Welter *et al.*, 2014). The two common designs are case-control studies, which look for associations between SNPs and disease, and population-based studies which focus on finding associations between SNPs and continuous traits (McCarthy *et al.*, 2008). The larger goal of these studies is to function as hypothesis-generating machines, resulting in sets of loci that require further analysis. Thus GWAS are an important first step in the gene identification process (Cantor *et al.*, 2010). The findings from these studies provide preliminary genetic information, which need additional analysis and follow-up experiments to be validated. However, many studies have found only a few common SNPs per trait, and these SNPs have generally low predictive power, explaining only a small percentage of the variance (Manolio *et al.*, 2009).

Often, SNPs are tested individually for association with the phenotype, using the Armitage Trend Test. Because genome-wide scans analyze hundreds of thousands or even millions of markers, the multiple testing issue is resolved by applying a stringent significance threshold - most commonly $5 * 10^{-8}$ (Panagiotou and Ioannidis, 2012) - to the p-values. This method is successful only if the study is well-powered, such that the associations are strong enough to pass the stringent threshold. However, even if that is the case, this type of analysis has several limitations, which have been addressed in the literature (Schork, 2001; Hoggart *et al.*, 2008; Li *et al.*, 2011; He and Lin, 2011; Rakitsch *et al.*, 2013). Here we focus on two of them. Firstly, single SNPs tend to have small effect sizes. We can increase the explanatory power by looking at the joint effect of multiple SNPs. Secondly, when we test a SNP individually, we ignore the effects of all other SNPs. If we analyze marginally two sufficiently correlated SNPs, out of which only one is causal for the disease, both may show an association. This leads to higher false positive rates.

Joint modeling of all SNPs is challenging. Since in most GWAS the number of SNPs is much larger than the number of samples, the data cannot be analyzed using standard multivariable approaches. An established method in the field is the GCTA (Genome-wide

*to whom correspondence should be addressed

Complex Trait Analysis), which is based on linear mixed models (Yang *et al.*, 2011, 2014) and enables some joint analysis of SNPs. It allows for statistical significance tests of single SNPs (as fixed effects) while all SNPs other than the considered single SNP are built into the model as a simultaneous random effect. We would classify the obtained statistical significance of the SNPs as a hybrid between marginal (with only one or a few SNPs as fixed effects) and joint (since all the SNPs are in the model) modeling. It also enables to assess the combined effect of all SNPs which quantifies the heritable component of phenotype variation explained jointly by all the genotyped SNPs (Yang *et al.*, 2010). Another solution to the high-dimensionality of the problem is the use of penalized regression, which constrains the magnitude of the regression coefficients, and allows them to be estimated. The two most widely used penalization methods are the Lasso (Tibshirani, 1996) and Ridge regression (Hoerl and Kennard, 1970). The Lasso penalizes the sum of the absolute values of the regression coefficients. It is a sparse estimator, meaning that it sets some regression coefficients to zero, while keeping others non-zero. Ridge regression penalizes the sum of squared regression coefficients, but it does not reduce the number of parameters in the model. In Abraham *et al.* (2013) it has been shown in the context of GWAS that penalization decreases the false positive rate and increases the probability of detecting the causal SNPs. There are several papers which consider a joint analysis. Methods which apply a penalized model include: the Bayesian Lasso (Li *et al.*, 2011), a two-stage procedure using single regression followed by a Lasso selection (Shi *et al.*, 2011), stability selection in the context of GWAS (Alexander and Lange, 2011), the so-called ISIS (Iterative Sure Independence Screening) combined with stability selection to select significant SNPs (He and Lin, 2011), a combination of Lasso and linear mixed models (Rakitsch *et al.*, 2013), Lasso for screening (Wu *et al.*, 2010) or ridge regression (Malo *et al.*, 2008). None of the proposals (Li *et al.*, 2011; He and Lin, 2011; Rakitsch *et al.*, 2013; Shi *et al.*, 2011; Alexander and Lange, 2011) compute p-values for SNPs. Wu *et al.* (2010) aims to control the type I error rate, the approaches using stability selection aim to control the expected number of false positive selections (Meinshausen and Bühlmann, 2010), while Shi *et al.* (2011) controls the False Discovery Rate (FDR).

Our goal is to construct valid p-values for SNPs in a (joint) multiple generalized linear model together with a computationally efficient and powerful way to address the issue of massive multiple statistical hypothesis testing. The problem is challenging due to the complex setting with hundreds of thousands of SNPs. Our method relies on a *hierarchical* procedure from Mandozzi and Bühlmann (2015) which we apply here for the first time to GWAS with very high-dimensional data-sets. It provides p-values for multiple (joint) regression modeling of SNPs in high-dimensional settings. We compute p-values not only for individual SNPs, but also for groups of SNPs. The idea is to adapt to the strength of the signal present in the data: if the signal is too weak or the SNPs exhibit too high correlation, we might still detect a significant group of SNPs, instead of single SNP markers. Additionally, we compute the explained variance for every such group in a high-dimensional generalized linear model.

We demonstrate our method on the WTCCC data (The Wellcome Trust Case Control Consortium, 2007), due to the fact that strong associations have been found for some phenotypes in this data set, and many of their findings have been replicated in subsequent studies. However, our method's advantages are also evident for phenotypes with weak associations: for biologically distant phenotypic traits, the goal is rather to find regions of the genome that are strongly associated with the phenotype. Our proposed method makes it statistically and computationally possible to assess the significance of the parameters in a multiple (generalized) linear model, for large scale GWAS problems with millions of SNP markers. The interpretation of the parameters in a multiple (generalized) linear model is markedly different from marginal association and also from GCTA (Yang *et al.*, 2011). In fact, under some assumptions, we can link the (joint) multiple linear model to causal inference (see Section 2.1). Thus, it as an important step to perform the statistical inference in a multiple generalized linear model.

2 METHODS

Consider the following setting and notation. There are n samples (e.g., persons in a study), and each of them is indexed with $i \in \{1, \ldots, n\}$. A response variable Y_i for the ith sample point (e.g., the ith person in the study) encodes the status of a phenotype of interest. For example the binary status of a disease with $Y_i \in \{0, 1\}$, the continuous value of a survival time with $Y_i \in \mathbb{R}^+$ or the continuous degree of an exposure or (log-) concentration with $Y_i \in \mathbb{R}$. The regressor X_i is a (long) $p \times 1$ vector which encodes the SNP profile for the ith sample point: $X_{i,j} \in \{0, 1, 2\}$ is the value of the jth SNP for sample point i, taking three possible values corresponding to the number of minor alleles per person. Typically, the number of SNPs (regressors) is $p \approx 10^6$, while the number of samples is at least one order of magnitude smaller. A model measuring multivariable association is introduced next.

2.1 Generalized linear models

A well-established model for relating the phenotype (response variable) and the SNPs (regressors) is a generalized linear model (McCullagh and Nelder, 1989).

The easiest form thereof is a linear model for continuous ($\mathbb{R}$-valued) responses:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i \ (i = 1, \ldots, n), \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed noise terms with expectation $\mathbb{E}[\varepsilon_i] = 0$, finite variance and which are uncorrelated with the regressors $X_{i,j}$.

For binary responses with $Y_i \in \{0, 1\}$, representing case (= 1) or control (= 0), we consider a logistic regression model:

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i), \\ \pi_i &= P(Y_i = 1 | X_i, \beta) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, (i = 1, \ldots, n), \\ \ln(\frac{\pi_i}{1 - \pi_i}) &= \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} \end{aligned} \tag{2}$$

Here, π_i represents the probability of individual i having a case status given its SNPs X_i. There is no additional noise term and the stochastic nature of the model comes from the probability π_i.

In both models, β_0 denotes the intercept, and the coefficients β_j are the (logistic) regression coefficients which measure the association of the jth SNP with the response. Such models, which take into account all SNPs, have two features. First, the (generalized) regression coefficients have the following (well-known) interpretation: β_j measures the association effect of

$X_{i,j}$ on Y_i which is not explained by all other variables $\{X_{i,k};\ k \neq j\}$. Thus, a large β_j, in absolute value, has the very powerful interpretation that SNP j has a strong association to the phenotype *given all other SNPs* or *controlling for all other SNPs*. This is in sharp contrast to marginal correlation between SNP j and the phenotype Y which can easily be of spurious nature and caused by another SNP k having a strong correlation with the phenotype and with SNP j.

Furthermore, the regression models are predictive in the sense that for a new sample point (e.g. person) with a given SNP profile X_{new} we obtain a prediction for the corresponding phenotype (e.g. disease status) $\mathbb{E}[Y_{\text{new}}|X_{\text{new}}] = \beta_0 + \sum_{j=1}^{p} \beta_j X_{\text{new},j}$ or $\mathbb{P}[Y_{\text{new}} = 1|X_{\text{new}}] = \eta_{\text{new}}$, where $\eta_{\text{new}} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{\text{new},j}$. Note that this prediction is likely to be more informative or precise than a prediction that is based merely on marginal correlations because the general linear model applied here enables us to use the *whole new SNP profile*, and not just single SNPs, for predictive purposes.

Our main goal is to infer statistical significance of a single SNP or of a possibly large group of correlated SNPs for a given phenotype. More precisely, we aim for p-values, adjusted for multiple testing, when testing the following hypotheses:

for single SNP j

$$H_{0,j} : \beta_j = 0 \text{ versus } H_{A,j} : \beta_j \neq 0, \tag{3}$$

or for a group $G \subseteq \{1, \ldots, p\}$ of SNPs

$$H_{0,G} : \beta_j = 0 \text{ for all } j \in G$$

$$\text{versus} \quad H_{A,G} : \text{at least for one } j \in G \text{ we have that } \beta_j \neq 0. \tag{4}$$

The obtained p-values are with respect to a regression model and hence, they share the interpretation with the regression parameters described above. In particular, they are markedly different from a marginal or linear mixed model approach: the differences are also illustrated in simulation studies in Section 3.1.

A link to causal inference. If we assume (i) that the model is correct and that beyond the measured SNPs there are no hidden confounding variables - a condition that might be somewhat less problematic when having a million or more SNP markers - and (ii) that the causes point from the SNPs to the phenotype Y, the parameters β_j $(j = 1, \ldots, p)$ can be given a causal interpretation. This link to causal inference shows again that a (joint) multiple regression model is very different from a marginal model. In a structural equation model the assumption that the causes point from the SNPs to the phenotype Y means that the arrows in a directed acyclic graph, that encode the causal influence diagram, point to Y and never point away from Y, i.e., Y is childless. Such an assumption says that some SNPs might be the cause for a phenotype, but the phenotype cannot be a cause for the SNPs, which seems a very reasonable assumption. Under these conditions, the following holds: if $\beta_j \neq 0$, then there must be a directed edge in the causal influence diagram of a linear structural equation model from SNP j to the phenotype Y with non-zero edge weight, i.e., there exists a non-zero direct causal effect from SNP j to the phenotype Y. This statement is not true with marginal associations (i.e., if SNP j is only marginally associated with Y) since adjusting for all other SNPs (different from SNP j) is crucial for causal statements. The details are given in Proposition S1.1 in the Supplementary Material Section S1.

2.2 The challenge of high-dimensionality

The difficulty with a regression type analysis is the sheer high-dimensionality of the problem. The number of SNPs $p \approx 10^6$ is massively larger than sample size n, which is at least one order of magnitude smaller. In such scenarios, standard statistical inference methods fail. Recent progress based on new methods such as multiple sample splitting, has allowed us to obtain statistical significance measures for regression parameters β_j (Meinshausen *et al.*, 2009; Bühlmann, 2013; Zhang and Zhang, 2014, cf.) or groups thereof (Mandozzi and Bühlmann, 2015). We rely here on this method (Mandozzi and Bühlmann, 2015), which shows reliable performance over a wide range of simulation settings (Dezeure *et al.*, 2015), and enjoys the property of being computationally vastly more efficient than procedures which operate on the entire data-set. We extend the procedure from Mandozzi and Bühlmann (2015) from linear to logistic regression, and we show here for the first time how it performs for extremely high-dimensional GWAS data. The entire statistical procedure is schematically summarized in Figure 1.

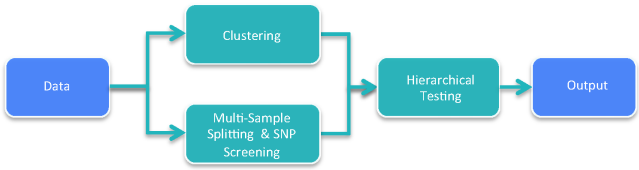

Fig. 1. Schematic overview of the method. "Clustering" refers to the step of hierarchically clustering the SNPs. SNPs on different chromosomes are clustered separately, after which the 22 clusters are joined into one final cluster containing all SNPs. " Multi-Sample Splitting and SNP Screening" stands for the SNP selection in steps 1 and 2 of the method described in Section 2.4.2. These selected SNPs are used to compute the p-values. Finally, the last step of the method - "Hierarchical Testing" - uses the selected SNPs to test groups of SNPs and eventually single SNPs. This testing is done hierarchically, on the cluster previously constructed. The output of the method consists of significant groups , or single SNPs, along with their p-values, that are adjusted for multiple testing.

In view of the high-dimensional nature of GWAS, it is rather unlikely to detect single SNPs which are significant when *controlling for all other SNPs*. Thus, it is a-priori more likely to detect (large) significant groups of SNPs with respect to the group hypotheses $H_{0,G}$ in a regression model. The construction of such groups is achieved by clustering the SNPs, as explained next.

2.3 Clustering

Our goal is to perform significance testing on single SNPs (the hypotheses $H_{0,j}$) as well as arbitrarily large groups of SNPs (the hypotheses $H_{0,G}$). We do this *hierarchically* since this allows for powerful multiple testing adjustment as well as for efficient computation (see Section 2.4).

We first discuss the hierarchical clustering of SNPs. The hierarchy can be constructed in different ways. One option is to use specific domain knowledge to group the SNPs, for instance by clustering them first into genes, and then into functional pathways. Another option is to use standard hierarchical clustering methods which rely on a distance measure between the SNPs.

Here we adopt the second approach, which is similar to the construction of haplotype maps (Barrett *et al.*, 2005). We use hierarchical clustering with average linkage (Jain and Dubes, 1988) which can be represented as a cluster tree, denoted by $\mathcal{T}$. The method requires a distance or dissimilarity measure between SNPs. We consider the distance between two SNPs as one minus their linkage disequilibrium (LD) value, where LD refers to the statistical dependency of the DNA content at nearby locations of the chromosome. One of the most common measures of LD is the square of the Pearson correlation coefficient (Hill and Robertson, 1968), which quantifies the linear dependence between two loci. Thus, two SNPs will have an LD equal to one if they are perfectly correlated, or an LD equal to zero if they are uncorrelated. Since LD has a tendency to decay with the distance of the studied loci, close-by SNPs are typically in high LD. This means that SNPs belonging to the same gene, or more generally, neighboring SNPs will end up in the same cluster. Often, LD is studied within each chromosome separately.

Therefore, we construct separate cluster trees for each chromosome [1] , and we then join these into one tree $\mathcal{T}$ which contains all the SNPs in the study, as shown in Figure 2.

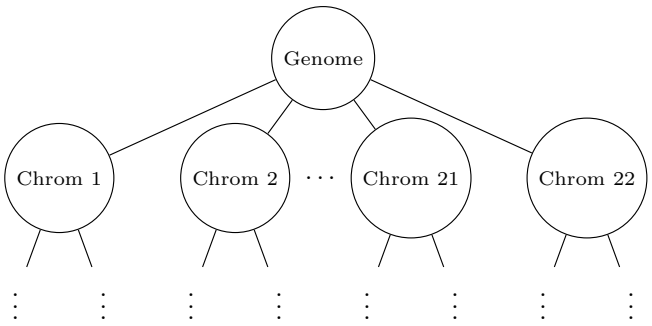

Fig. 2. The final cluster tree. The SNPs are first partitioned into chromosomes, and then a cluster tree is built for each chromosome separately using hierarchical clustering with average linkage. The hierarchical clusters of SNPs within chromosomes are not shown due to their size.

2.4 Statistical Significance Testing

A cluster, as described in Section 2.3, is denoted by the generic letter G which encodes a subset of $\{1, \ldots, p\}$ of single SNPs. We explain here how to test a null-hypothesis for a group $H_{0,G}$ in (4) or for a single SNP $H_{0,j}$ in (3).

2.4.1 *Hierarchical inference*

In section 2.4.2 we will show how one can construct valid p-values for the hypotheses $H_{0,j}$ and $H_{0,G}$. On the basis of valid p-values, our hierarchical approach proceeds as follows:

1. Test the global hypothesis $H_{0,G_{\text{global}}}$ where $G_{\text{global}} = \{1, \ldots, p\}$: that is, we test whether all SNPs have corresponding (generalized) regression coefficients equal to zero or alternatively, whether there is at least one SNP which has a non-zero regression coefficient. If we can reject this global hypothesis, we go to the next step.
2. Test the hypotheses $H_{0,G_1}, \ldots, H_{0,G_{22}}$ where G_k contains all the SNPs from chromosome k. For those chromosomes k where H_{0,G_k} can be rejected, we go to the next step.
3. Test hierarchically the groups G which correspond to chromosomes k where H_{0,G_k} was previously rejected. Consider first the largest groups and then proceed hierarchically (down the cluster tree) to smaller groups until a hypothesis $H_{0,G}$ cannot be rejected anymore or the level of single SNPs is reached.
4. The output is a collection of groups $G_{\text{final},1}, \ldots, G_{\text{final},m}$ where $H_{0,G_{\text{final},k}}$ is rejected ($k = 1, \ldots, m$) and all subgroups of $G_{\text{final},k}$ ($k = 1, \ldots, m$) downwards in the cluster tree are not significant anymore.

In such a hierarchical testing procedure, which belongs to the scheme of sequential multiple hypothesis testing, the multiple testing adjustment is resolution dependent. To guarantee that the familywise error, i.e., the probability for at least one false rejection of the hypotheses among the multiple tests, is smaller than or equal to α for some pre-specified $0 < \alpha < 1$, e.g., $\alpha = 0.05$, the hypothesis tests must be performed at different significance levels, depending on where one is in the hierarchy. The more we descend in the hierarchy, the more the multiple testing adjustment increases because we do more tests. It is important to keep in mind that even though the procedure controls the type I error simultaneously over all levels of the hierarchy, the adjustment for larger clusters does not depend on whether one will test their subclusters or not. While there is an ordering of the clusters, due to the nature of the hierarchical clustering procedure, and testing of subclusters stops once the null hypothesis of the parent cluster is accepted, the adjustment applied to the p-value of any cluster does not depend on the number of tests that have already been performed, but only (essentially) on the size of that particular cluster, see Section 2.4.2. The details have been developed by Meinshausen (2008). In other words: the final output of the method are p-values for significant groups $G_{\text{final},1}, \ldots, G_{\text{final},m}$ with the interpretation, that these p-values control the familywise error rate for multiple testing.[2] Furthermore, due to the hierarchical structure of the procedure, we can massively reduce the number of computations: if the final clusters or groups $G_{\text{final},k}$ are relatively high up in the hierarchy of the cluster tree, we only need to compute relatively few hypothesis tests.

2.4.2 *Construction of the p-values*

The hierarchical inference procedure above assumes that one has a method that constructs p-values which are valid.[3]

Due to the high-dimensionality with $p \gg n$, obtaining a p-value for the hypotheses $H_{0,j}$ or $H_{0,G}$ in (3) or (4) is a non-trivial problem. We rely here on a multiple sample splitting approach from Meinshausen *et al.* (2009), and we follow exactly the method from Mandozzi and Bühlmann (2015). The idea is as follows. For $b = 1, \ldots, B$ repetitions:

1. Randomly partition the n samples into two parts, say $N_{in}^{(b)}$ and $N_{out}^{(b)}$.
2. Using a variable selection procedure such as the (logistic) Lasso (Tibshirani, 1996; Friedman *et al.*, 2010), select regressors (SNPs) based on data from the first half-sample $N_{in}^{(b)}$. Denote the selected regressors by $\hat{S}^{(b)} \subseteq \{1, \ldots, p\}$. Because a Lasso estimated model has cardinality smaller or equal to $\min(n, p)$, the number of selected variables $|\hat{S}^{(b)}| < n/2$ will be smaller than half of the sample size . We choose to select the first $n/6$ SNPs that enter the Lasso path. This ensures that we have enough regressors for computing p-values.
3. Based on data from the second half-sample $N_{out}^{(b)}$, use classical p-value constructions in a linear or generalized linear model with the selected regressors (SNPs) from $\hat{S}^{(b)}$ in the previous step. The construction of a p-value of a cluster G is done in the following manner: we intersect the hierarchy $\mathcal{T}$ constructed in Section 2.3 (using hierarchical clustering) with $\hat{S}^{(b)}$, obtaining an induced hierarchy with root node $\hat{S}^{(b)}$. The testing is then applied on this induced hierarchy. Finally we assign the p-value to the entire cluster G, although we have only used the variables in $G \cap \hat{S}^{(b)}$.

$$p^{G,(b)} = \begin{cases} p_{out}^{G \cap \hat{S}^{(b)}} \text{ based on } Y_{N_{out}^{(b)}}, X_{N_{out}^{(b)}}, & \text{if } G \cap \hat{S}^{(b)} \neq \emptyset \\ 1, \text{ if } G \cap \hat{S}^{(b)} = \emptyset, \end{cases} \tag{5}$$

where $p_{out}^{G'}$ is the p-value for $H_{0,G'}$ based on data from $N_{out}^{(b)}$ ($G' \subseteq \{1, \ldots, p\}$). For a cluster $G \in \mathcal{T}$, the multiplicity adjusted p-value is

[1] In addition to providing a biological interpretation, clustering each chromosome separately results in substantial computational gains for problems with $p \approx 10^6$ SNPs.

[2] If we collect all groups with a p-value smaller or equal to α, then the probability for making one or more false rejections among all considered tests is less or equal to α.

[3] A p-value P is valid for a null-hypothesis H_0 if $\mathbb{P}_{H_0}[P \leq \alpha] \leq \alpha$ for any $0 < \alpha < 1$, where $\mathbb{P}_{H_0}$ denotes the probability assuming that H_0 is true.

defined as:

$$p_{adj}^{G,(b)} = \min(p^{G,(b)} \frac{|\hat{S}^{(b)}|}{|G \cap \hat{S}^{(b)}|}, 1) \quad (6)$$

if $G \cap \hat{S}^{(b)} \neq \emptyset$ and $p_{adj}^{G,(b)} = 1$ otherwise.

4. Repeat steps 1-3 B times (with e.g. $B = 100$) and aggregate the B p-values (separately for every hypothesis). The aggregated p-value of any cluster G is computed by considering its empirical quantile:

$$P^G = \min\{1, (1 - \log \gamma_{min}) \inf_{\gamma \in (\gamma_{min}, 1)} Q^G(\gamma)\} \quad (7)$$

where $Q^G(\gamma) = \min\{1, q_\gamma(\{p_{adj}^{G,(b)}/\gamma; b = 1, ..., B\})\}, \gamma \in (0, 1)$, $\gamma_{min} = 0.05$ and $q_\gamma(\cdot)$ is the empirical γ-quantile function. Finally, the hierarchically adjusted p-value of a cluster G is:

$$P_h^G = \max_{D \in \mathcal{T}: G \subseteq D} P^G \quad (8)$$

The sample splitting in step 1 is made to avoid being over-optimistic when performing variable selection and p-value construction on the same data-set. The repeated sample splitting in step 4 helps to achieve much more reliable results which do not depend in a sensitive way on how we split the sample (Meinshausen *et al.*, 2009, cf.). More details about the assumptions which guarantee control of the familywise error rate are provided in Supplementary Material Section S2.

This multi-sample splitting method is computationally fast since Lasso in step 2 is rather cheap to perform and step 3 requires classical p-value computations in low-dimensional models with fewer than n regressors only. In terms of accuracy for type I error control, i.e., avoiding false rejections of hypotheses, the multi-sample splitting approach has been found very reliable in extensive simulations relative to other methods. This reliability comes at the price of being slightly inferior in terms of power to detect true underlying positive findings (Dezeure *et al.*, 2015), see also Mandozzi and Bühlmann (2015). However, this slightly more conservative scheme has the advantage of limiting false positives.

3 RESULTS

3.1 Simulation studies

We used the WTCCC Crohn's disease genotype data to create semi-synthetic datasets. To generate the new genotype matrix, we kept all the samples ($n = 4682$), but selected a block of 500 consecutive SNPs from each of the 22 autosomal chromosomes, having in total 11000 SNPs. The phenotype data was generated from a logistic regression model with the probability of having the disease as the dependent variable and 10 causal SNPs as the independent variables. We considered 3 designs for choosing the causal SNPs:

1. Randomly select a set of 10 consecutive SNPs from chromosome 1. The regression coefficients are sampled with replacement from the set {-2, -1.75, -1.5, -1.25, -1, 1, 1.25, 1.5, 1.75, 2}.
2. Randomly select a set of 5 consecutive SNPs from chromosome 1 and 5 consecutive SNPs from chromosome 2. The regression coefficients are sampled with replacement from the set {-1, -0.75, -0.5, 0.5, 0.75, 1}.
3. Randomly select a set of 10 non-consecutive SNPs from chromosome 1. The regression coefficients are sampled with replacement from the set {-2, -1.75, -1.5, -1.25, -1, 1, 1.25, 1.5, 1.75, 2}.

To ensure that the number of cases and controls are not too different, we required the ratio between cases and controls to be within the interval $[0.67, 1.5]$. We kept the genotype matrix constant, and generated 100 simulation runs for each design, using new coefficients for every simulation run.

We chose to compare our method to three other algorithms. One is the classic bivariate testing, implemented in PLINK (Purcell *et al.*, 2007). The other two are mixed model approaches: the FaST-LMM (Lippert *et al.*, 2011) and the GCTA (Yang *et al.*, 2011) algorithms. Both calculate a genetic relationship matrix (GRM) to control for the effect of the other SNPs. There are many ways of computing the GRM. One can use all the SNPs, or just a particular subset. An approach that is computationally efficient is leave-one-chromosome-out (LOCO) (Yang *et al.*, 2014). With this option when one tests the SNPs in a particular chromosome, the SNPs from all the other chromosomes besides the one being tested are used to compute the GRM. The main difference between mixed models and our method is the way in which the effects of other SNPs are modeled. While the mixed model uses a random component to account for all the other SNPs, our method considers each SNP as a fixed effect, and includes all of them in the model.

Our goal was to assess how good these methods are at detecting the causal variants (which are known for simulated data), while limiting the number of false positives: SNPs that are not truly causal (but perhaps correlated with the causal variants). Assessing the performance of the methods was done by considering several criteria. The first is the FWER, which we expect to be controlled at level α. This is equivalent to expecting $100*\alpha$ false discoveries, when performing 100 simulations. As a less conservative criterion, we also consider the k-FWER, a generalized version of the FWER. The k-FWER is defined as $P\{V \geq k\}$, where V is the total number of false rejections. For our simulations, we are interested in the value of k, under which the k-FWER is controlled at level $\alpha = 0.05$. In the case of our method, a rejection is considered false only if the cluster does not contain any of the true causal SNPs. The third assessment criteria is the power of the method. Also here we consider two variants. The first is a "naive" version, which considers all findings where the true causal SNP is present in a cluster, irrespective of the cluster size. The second metric penalizes the size of the group with respect to the causal variants. This is computed in the following way:

$$POW_{adaptive} = \frac{1}{|S_0|} \sum_{G \in G_{sign}} \frac{|S_0 \cap G|}{|G|}, \quad (9)$$

where G_{sign} is the set of groups declared significant by our method, and S_0 is the set of true causal SNPs. For the three comparison methods, we declare significant SNPs that have a p-value below $5 * 10^{-8}$, and compute the power, FWER and k-FWER using this set.

The results shown in Table 1 are in line with our expectations. In the first 2 designs, our method has a lower power compared to the other three methods that behave almost identically. The cost of a larger power is however a significant increase in the number of false positives. While our method fails to control the FWER due to the very complex correlation structure in the data, it does however control the 2-FWER at level α. This means that the probability of making more than 2 false rejections is below α. In comparison, the other three methods do on average at least one order of magnitude more false rejections. This has to do with the fact that they infer

Table 1. Simulation results. Comparison of four methods for three different scenarios. FWER: Familywise error rate; k: value of k such that k-FWER ≤ 0.05; POW: power; $\text{POW}_{adaptive}$: adaptive power.

Design	Method	FWER	k	POW	$\text{POW}_{adaptive}$
1	hierGWAS	0.14	2	0.70	0.63
1	PLINK	1	44	0.89	
1	GCTA	1	44	0.89	
1	FaST-LMM	1	44	0.89	
2	hierGWAS	0.29	2	0.72	0.66
2	PLINK	1	81	0.87	
2	GCTA	1	93	0.89	
2	FaST-LMM	1	93	0.89	
3	hierGWAS	0.56	3	0.94	0.85
3	PLINK	1	130	0.94	
3	GCTA	1	131	0.94	
3	FaST-LMM	1	130	0.94	

marginal associations, or associations which are partially adjusted by random effects: many significant findings are spurious because of high correlation between some of the SNPs. The lower power of our hierGWAS procedure can be explained by the fact that it aims to infer associations which are adjusted for other SNPs: our method would detect some of them individually, some of them as groups, and if the correlation among some SNPs is too strong (or the signal too weak), it would miss some. This becomes apparent also when we consider the two power measures: $\text{POW}_{adaptive}$ is always smaller than POW, because some of the causal SNPs will be grouped into clusters, due to their correlation structure. In the case of the third design, our method has the same power as the other three to detect the causal SNPs. If we assume that this is close in spirit to the real life case of having one causal SNP surrounded by many others that are in LD with the causal one, our method has the power to detect it just as well as the marginal methods, while providing a greatly reduced set of false positives, and a much stronger interpretation of the findings. While the FWER is the highest in the third design, due to the fact that there are many more confounders around each SNP compared to the other two designs, our method still performs much better, by controlling the 3-FWER at level α. We note that GCTA and FaST-LMM have a slight disadvantage regarding the control of false positives, because we used the LOCO approach to compute the GRM. However, since this is an established approach, and the strategy of eliminating only the SNP being tested, and using all other SNPs to construct the GRM is computationally infeasible (Yang *et al.*, 2014), we believe that this situation reflects reality.

3.2 WTCCC data

We validate our method on data from The Wellcome Trust Case Control Consortium (2007). The Wellcome Trust Case Control Consortium study used 3000 subjects and 2000 shared controls from the British population to examine 7 major diseases: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). The subjects were genotyped using the Affymetrix GeneChip 500K Mapping Array Set. Though The Wellcome Trust Case Control Consortium (2007) reported all SNPs with a p-value $< 5 * 10^{-4}$, the threshold for strong association was set to $5 * 10^{-7}$. Using the standard marginal analysis, the WTCCC study identified 21 new SNPs strongly associated to the phenotype. For BD rs420259 on chromosome 16, for CAD rs1333049 on chromosome 9, for CD, the WTCCC study identified 9 SNPs strongly associated to the phenotype: rs11805303 on chromosome 1, rs10210302 on chromosome 2, rs9858542 on chromosome 3, rs17234657 and rs1000113 on chromosome 5, rs10761659 and rs10883365 on chromosome 10, rs17221417 on chromosome 16 and finally rs2542151 on chromosome 18. 2 SNPs were found for RA: rs6679677 on chromosome 1 and rs6457617 on chromosome 6. T1D was strongly associated to 5 SNPs: rs6679677 on chromosome 1, rs9272346 on chromosome 6, rs11171739 and rs17696736 on chromosome 12 and rs12708716 on chromosome 16. Finally, for T2D 3 associations were found: rs9465871 on chromosome 6, rs4506565 on chromosome 10 and rs9939609 on chromosome 16. For HT, the WTCCC did not find any SNP strongly associated to the phenotype.

Before applying our analysis, we have preprocessed the data, by excluding some SNPs and samples, as well as imputing the missing SNPs. Details about this procedure are given in the Supplementary Material Section S4.1.

The output of our method is a list of SNP groups of different sizes. These represent the smallest jointly significant groups in the hierarchical tree of SNPs. We create a distinction between small (< 10 SNPs) and large groups, and present the corresponding results separately. The number 10 is somewhat arbitrary and determined by notational simplicity to list at most 10 SNPs per group. We identified small groups for 5 of the 7 diseases, and present them below. Large groups have been identified for all of the diseases, however we chose to present in detail the results for BD only. We chose BD because it is the disease for which the WTCCC found a single strongly associated SNP, which we did not identify using our method. The large groups for the other 6 diseases are detailed in the Supplementary Material Section S4. It is important to note that these groups are not overlapping. For example, if in a specific chromosome we find a small group of 4 SNPs, as well as 2 large groups both containing thousands of SNPs, these 3 groups do not share common SNPs and they belong to different regions of the chromosome. This happens because our method finds the smallest group of SNPs for which the null hypothesis can be rejected. Such a result means that one region of the chromosome exhibits a strong signal, while there are other regions exhibiting weaker signal. Thus, the size of the group reflects the strength of associations: the weaker these associations, the larger the significant groups.

Tables 2 and 3 report on individual SNPs or small clusters of SNPs selected by our method for the seven diseases we analyzed. We found a total of 20 such clusters, out of which 16 are individual SNPs.

12 out of the 20 clusters contain at least one SNP that was found to be strongly associated to the phenotype in the original WTCCC study. The remaining 8 clusters contain SNPs that are either in LD with the ones identified by the WTCCC, belong to the same gene or genomic region, or have been identified as having a significant effect in other studies. While it is informative to see if our findings have been previously reported in other studies, it is important to remember the distinction in terms of interpretation. Our method makes the significance of previous findings much stronger, because it does not simply compute the marginal correlation, but it instead

Table 2. List of small significant groups of SNPs selected by our method for coronary artery disease, Crohn's disease and rheumatoid arthritis. [a] The disease identifier for which the SNP group was selected. [b] The smallest groups of SNPs whose null hypothesis was rejected. The SNPs in this group are jointly significant. rsIDs of SNPs from dbSNP. [c] The chromosome to which the SNPs in the group belong. [d] The gene to which the SNPs in the group belong, if any. Gene symbol from Entrez Gene. [e] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER). [f] The variance explained by the group of SNPs.

Dis[a]	Significant group of SNPs[b]	Chr[c]	Gene[d]	P-value[e]	R^{2f}
CAD	rs1333049	9	intergenic	$1.7 * 10^{-3}$	0.013
CD	rs11805303, rs2201841, rs11209033, rs12141431, rs12119179	1	IL23R	$4.5 * 10^{-2}$	0.014
CD	rs10210302	2	ATG16L1	$4.6 * 10^{-5}$	0.014
CD	rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734, rs16869934	5	intergenic	$2.7 * 10^{-3}$	0.016
CD	rs10883371	10	LINC01475, NKX2-3	$2.4 * 10^{-2}$	0.004
CD	rs10761659	10	ZNF365	$1.5 * 10^{-2}$	0.007
CD	rs2076756	16	NOD2	$1.3 * 10^{-3}$	0.017
CD	rs2542151	18	intergenic	$1.5 * 10^{-2}$	0.005
RA	rs6679677	1	PHTF1	$5.9*10^{-11}$	0.031
RA	rs9272346	6	HLA-DQA1	$1.4 * 10^{-6}$	0.017

Table 3. List of small significant groups of SNPs selected by our method for type 1 diabetes and type 2 diabetes. [a] The disease identifier for which the SNP group was selected. [b] The smallest groups of SNPs whose null hypothesis was rejected. The SNPs in this group are jointly significant. rsIDs of SNPs from dbSNP. [c] The chromosome to which the SNPs in the group belong. [d] The gene to which the SNPs in the group belong, if any. Gene symbol from Entrez Gene. [e] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER). [f] The variance explained by the group of SNPs.

Dis[a]	Significant group of SNPs[b]	Chr[c]	Gene[d]	P-value[e]	R^{2f}
T1D	rs6679677	1	PHTF1	$3.6*10^{-11}$	0.03
T1D	rs17388568	4	ADAD1	$2.7 * 10^{-2}$	0.006
T1D	rs9272346	6	HLA-DQA1	$2.4 * 10^{-3}$	0.17
T1D	rs9272723	6	HLA-DQA1	$2.2 * 10^{-4}$	0.17
T1D	rs2523691	6	intergenic	$6.04 * 10^{-5}$	0.004
T1D	rs11171739	12	intergenic	$1.3 * 10^{-2}$	0.01
T1D	rs17696736	12	NAA25	$6.5 * 10^{-4}$	0.018
T1D	rs12924729	16	CLEC16A	$3.4 * 10^{-2}$	0.007
T2D	rs4074720, rs10787472, rs7077039, rs11196208, rs11196205, rs10885409, rs12243326, rs4132670, rs7901695, rs4506565	10	TCF7L2	$1.7 * 10^{-5}$	0.015
T2D	rs9926289, rs7193144, rs8050136, rs9939609	16	FTO	$4.7 * 10^{-2}$	0.007

tests whether the effect of a SNP or a group is still significant after we have taken into account the effect of all other SNPs. In the case of the small clusters, it restricts the confounders to a reduced number of SNPs that are either introns in the same gene, or in close proximity to each other. Besides these small significant groups, our method also identified larger groups. Again, it is important to keep in mind that these larger groups are not overlapping with the smaller ones, and they are in other regions of the chromosome. These clusters contain many of the SNPs that were identified to have moderate associations in the original study. Because of their size, they are given lower weight in terms of power, however, they reflect the assumption that these diseases are highly polygenic, and associations appear in many places throughout the genome. In the following we will describe in more detail our findings for each disease.

3.2.1 *Coronary artery disease (CAD)*

We replicated rs1333049, an intergenic SNP on chromosome 9, the only finding from the WTCCC study. Our result however has a much stronger interpretation compared to the original finding, because we control for all possible confounders. Thus, rs1333049 shows an association with the phenotype, even after taking into account the effects of all other SNPs.

3.2.2 *Crohn's disease (CD)*

On chromosome 1 we identified a small significant cluster of 5 SNPs: rs11805303, rs2201841, rs11209033, rs12141431 and rs12119179. Two of them: rs11805303 and rs2201841 are introns in the IL23R gene, while the last 3 SNPs are up to 22-kb downstream from IL23R. Though rs11805303 showed strong association in the WTCCC study (The Wellcome Trust Case Control Consortium, 2007), our result has a different and much stronger interpretation. Our finding is a group of 5 SNPs that are jointly significant, though none of them is significant individually. Because this significance results from a joint model of all SNPs, it means that our group is jointly significant while controlling for all other SNPs in the study. The interpretation of this finding is that we limit the set

of confounding SNPs to 4 other SNPs. When a SNP is declared significant by computing the marginal correlation, like in The Wellcome Trust Case Control Consortium (2007), the set of possible confounders that produce this correlation is the set of all other SNPs. Thus, if the correlation turns out to be a spurious one, there could be hundreds of other SNPs that produce it. In contrast, our method not only drastically reduces the number of confounders, but gives a small set of much more plausible ones, that are in a narrow region of the chromosome, often clustered around a gene. On chromosome 2 we identified an individual SNP, rs10210302, which showed strong association in the WTCCC paper (The Wellcome Trust Case Control Consortium, 2007). This SNP has by far the lowest p-value ($4.6 * 10^{-5}$) in CD, and it also explains a relatively large proportion of the variance attributed to chromosome 2: 0.014 compared to 0.05 explained by all the selected SNPs in the chromosome. On chromosome 5 we identified a group of 9 SNPs: rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734 and rs16869934. They are all intergenic, 85 kb apart and located in the 40.4M region of the chromosome. rs16869934 is 4kb downstream from the SNP rs17234657 showing strong association to CD in The Wellcome Trust Case Control Consortium (2007). On chromosome 10 we found 2 significant SNPs. The first is rs10883371, a 2-kb upstream variant both for LINC01475 and NKX2-3. rs10883365, found to be strongly associated to CD in the WTCCC study is a 2-kb upstream variant in LINC01475. Our second finding on chromosome 10 is rs10761659, a non-coding intergenic SNP mapping 14-kb telomeric to gene ZNF365 and was identified first by the WTCCC (The Wellcome Trust Case Control Consortium, 2007), followed by a meta-analysis (Franke *et al.*, 2010), and later a study of a southern european population by Julia *et al.* (2013). On chromosome 16, we found an individual SNP, rs2076756, which is an intron in NOD2. Interestingly, this SNP was not found to be significant in the original WTCCC study, while our approach shows that it is even significant when we control for all other SNPs. This SNP has been confirmed by several studies (Franke *et al.*, 2010; Julia *et al.*, 2013; Rioux *et al.*, 2007; Kenny *et al.*, 2012). Finally, on chromosome 18, we identified a single intergenic SNP: rs2542151. This finding was reported by The Wellcome Trust Case Control Consortium (2007), as well as by Parkes *et al.* (2007).

3.2.3 *Rheumatoid arthritis (RA)*

We identified two SNPs, both individually significant. The first, rs6679677 is located on chromosome 1 and is a 2-kb upstream variant in the PHTF1 gene. This finding was reported by The Wellcome Trust Case Control Consortium (2007). The second SNP, rs9272346, is located on chromosome 6 and is also a 2-kb upstream variant in the HLA-DQA1 gene. This SNP belongs to the MHC region, just like the WTCCC finding.

3.2.4 *Type 1 diabetes (T1D)*

8 individual SNPs were declared significant by our method. 5 of these are the 5 associations found in The Wellcome Trust Case Control Consortium (2007). These are: rs6679677, a 2-kb upstream variant in the PHTF1 gene on chromosome 1, rs9272346, a 2-kb upstream variant in the HLA-DQA1 gene on chromosome 6, rs11171739, an intergenic SNP on chromosome 12, rs17696736, an intron in the NAA25 gene on chromosome 12 and rs12924729, an intron in the CLEC16A gene on chromosome 16. Additionally, our method identified 3 new associations. Two of them are located on chromosome 6: rs9272723 is an intron in the HLA-DQA1 gene and rs2523691 is intergenic. The third new finding, rs17388568, is located on chromosome 4 and is an intron in the ADAD1 gene. It did not reach the genome wide significance threshold in the WTCCC study, however it showed moderate association with a p-value of $3 * 10^{-6}$. It also showed moderate association in an independent study by Plagnol *et al.* (2011).

3.2.5 *Type 2 diabetes (T2D)*

We identified two small SNP clusters, one on chromosome 10 and the other on chromosome 16. The first cluster contains 10 SNPs: rs4074720, rs10787472, rs7077039, rs11196208, rs11196205, rs10885409, rs12243326, rs4132670, rs7901695, rs4506565, all introns in the TCF7L2 gene, spanning a 62KB region. One of them, rs4506565, was originally identified by The Wellcome Trust Case Control Consortium (2007), while rs7901695 showed a significant association in a replication study by Zeggini *et al.* (2007). The second cluster is comprised of 4 SNPs: rs9926289, rs7193144, rs8050136, rs9939609, all introns in the FTO gene spanning 10KB. rs9939609 was significantly associated to the phenotype in The Wellcome Trust Case Control Consortium (2007). Additionally, rs8050136 was found to have strong significance in Zeggini *et al.* (2007) and Scott *et al.* (2007).

3.2.6 *Bipolar disorder (BD)*

For BD, the WTCCC identified only one SNP strongly associated to the phenotype: rs420259. While we did not identify it in a small group, it is present in the large group found to be significant on chromosome 16. Furthermore, as can be seen in Table 4, we found clusters in many of the chromosomes. Table 4 shows the group size, both in terms of number of SNPs, as well as in terms of percentage of the total SNPs in that particular chromosome. Additionally, we investigate whether the SNPs identified using the standard analysis with PLINK (Purcell *et al.*, 2007) map into our groups. The size of the group is in a way inversely proportional to the strength of associations. If a certain chromosome contains SNPs with large effects, we will be able to find them in very small clusters, or maybe even individually. If however the signal is weak, we can only identify larger regions. For example, on chromosomes 4,6,7,8,10 and 15 the signal is so weak that we can only report that the joint effect of all SNPs in these chromosomes is significant, but we cannot further localize the signal. On the other hand, on chromosome 3 we were able to identify a much smaller group containing only 6 % of the SNPs.

Our method returns the smallest number of SNPs for which we can find a significant effect, while controlling for all other SNPs. The fact that we cannot disaggregate the signal to small clusters, or single SNPs does not mean that genetics plays no role in BD, but rather that the signal is very dispersed and the effect sizes are very small. This explains why our groups are so large, and why we cannot attribute the signal to narrower regions. Figure 3 shows the variance in bipolar disorder explained by the SNPs on individual chromosomes. We only consider the SNPs selected by the Lasso, in step 2 of Section 2.4.2, as these SNPs are a proxy for the truly

relevant SNPs. The total variance explained by all the SNPs is 0.5, and Figure 3 describes how this variation is distributed across the chromosomes. The fitted line corresponds to a linear model, where the predictor is the chromosome length, and the response is the explained variance.The plot gives weight to our previous statement, as the variance is homogeneously distributed across the chromosomes. The plot shows an excellent fit ($R^2 \approx 0.91$), meaning that the length of a chromosome is a very good predictor for the amount of variance that a particular chromosome explains.

3.2.7 Hypertension (HT)

Hypertension was the only disease where The Wellcome Trust Case Control Consortium (2007) did not find any strongly associated SNPs. We also haven't found small clusters or individual SNPs, but we did find larger clusters on 13 of the chromosomes. The results are shown in the Supplementary Material Section S4.7.

Table 4. List of large significant groups of SNPs selected by our method for bipolar disorder. [a] The size of the SNP group is the number of SNPs that belong to the group. In parenthesis: size as percentage of total genotyped SNPs on the chromosome. [b] The chromosome to which the SNPs in the group belong. [c] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER). [d] The variance explained by the group of SNPs. [e] We counted the number of SNPs with p-values $< 5*10^{-4}$ identified using PLINK (Purcell *et al.*, 2007). We looked at how many of those SNPs are present in the groups selected by our method. The numbers refer to the SNPs on individual chromosomes.

Size of significant SNP group[a]	Chr[b]	P-value[c]	R^{2d}	Hits[e]
6695 (22 %)	1	0.027	0.014	3 out of 10
12134 (40 %)	1	0.047	0.019	5 out of 10
14451 (45 %)	2	0.016	0.022	8 out of 18
7338 (23 %)	2	0.036	0.014	9 out of 18
1649 (6 %)	3	0.021	0.009	6 out of 15
24832 (100 %)	4	0.008	0.029	5 out of 5
14040 (55 %)	5	0.030	0.018	1 out of 5
24193 (100 %)	6	0.041	0.026	7 out of 7
20643 (100 %)	7	0.013	0.028	5 out of 5
21594 (100 %)	8	0.027	0.023	6 out of 6
11929 (65 %)	9	0.009	0.020	10 out of 12
22517 (100 %)	10	0.021	0.024	6 out of 6
15269 (77 %)	12	0.038	0.016	1 out of 2
4389 (36 %)	14	0.048	0.012	3 out of 11
11055 (100 %)	15	0.032	0.017	4 out of 4
10382 (88 %)	16	0.047	0.018	16 out of 16

4 DISCUSSION

We have presented a new method for assigning statistical significance in GWAS. Our approach goes beyond the bivariate testing of individual SNPs that looks only at marginal associations. Instead, we use a multivariable approach which includes all the SNPs and controls the familywise error rate. We propose to assign p-values in a hierarchical manner: first for chromosomes, and then

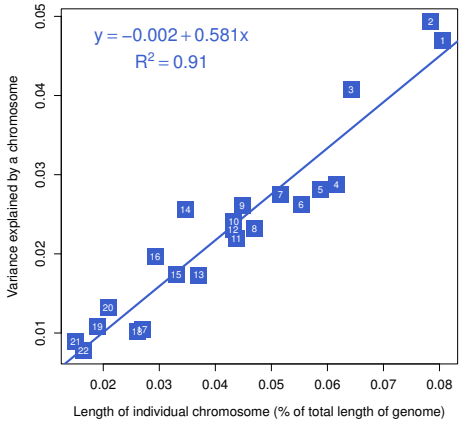

Fig. 3. Variance in bipolar disorder that is explained by individual chromosomes. The variance on the vertical axis is given by the R^2 value of all the selected SNPs in a chromosome, as described in the Supplementary Material Section S3. The total variance explained by all the selected SNPs on all the chromosomes is 0.5.

in a top-down fashion from larger to smaller groups of SNPs. Such an approach addresses several issues. First, since regression parameters of an individual SNP are typically very small, due to their interpretation and meaning in the model, it is much more likely to detect significant groups of SNPs. Second, because we proceed hierarchically, the problem of multiple testing is much less severe than for the classical one-SNP-at-a-time approach: roughly speaking, one has to adjust only for the number of tests which are considered, and this number is typically much smaller than the entire number of SNPs in the study. Our method is data-driven in the sense that its resolution for the groups of SNPs depends on the strength of the signal present in the data: how much we proceed in the hierarchy and refine the clusters of SNPs depends on how strong the associations are. If the signal is strong and well-localized, we find small clusters or individual SNPs, whereas if the signal is weak, we identify larger regions.

We demonstrate our method on the WTCCC data (The Wellcome Trust Case Control Consortium, 2007), where we analyze the seven diseases. Though it is interesting to conceptually validate our findings by comparing them with a measure of marginal association, our method is different and allows for a more powerful interpretation of the findings than testing only marginal association between a SNP and the phenotype. This is because we test whether or not SNPs in a cluster carry any additional information about the phenotype, beyond that available through all the other SNPs. That is, we adjust for the effect of all other SNPs that are not part of this cluster, which translates to a very strong interpretation of the significant clusters. This can be related to causal statements when making additional assumptions (see last paragraph in Section 2.1). Due to the fact that we control for all other SNPs, often we can reduce the number of possible confounders from hundreds or thousands of SNPs to less than 10. Moreover, our possible confounders are desirable candidates, as they are usually part of the same functional unit. This is a favorable outcome because in most cases it is unclear which is the causal SNP, and in many contexts the gene might be the more meaningful biological unit. Even for phenotypes with weaker and more dispersed signal, such as BD and HT, we could still identify

larger regions. While these clusters might be too large to identify specific genes, we can still gain insights into the joint influence of all selected SNPs, or the distribution of the variance across the chromosomes. This case is the one which motivated our approach. For distant, non-disease related phenotypes it is perhaps more useful to identify the chromosomes, or the regions that drive the signal, and their contribution to the total explained variance. In such cases identifying single SNPs is most likely impossible, and due to their low predictive power, not very useful.

It is difficult to directly compare our results to other marginal methods because we assign significance with respect to a generalized multiple regression parameter, and not only for individual but also for groups of SNPs. Nevertheless, we performed a small simulation study in which we compared the results for our method to the standard marginal approach, as well as two mixed model algorithms. The findings were in line with our expectations: while our method had slightly reduced power in two of the settings, it compensated by producing a significantly reduced number of false positive selections. In the third design our method had the same power as the mixed model and marginal testing approaches, while still having a superior control of the false positives.

One direction for improving the method would be to change the clustering, which could be performed through the use of more in-depth biological knowledge. For instance, if we would cluster the SNPs into genes, and then into pathways, even for a weak signal we would potentially identify larger pathways, which would be useful in terms of biological meaning.

Funding: This work was supported by the Advanced Investigator European Research Council Grant on the "Foundations of Economic Preferences" [295642 to L.B. and E.F.]; and the German National Science Foundation [SCHU 2828/2-1 to D.S.].

Conflict of interest : none declared.

REFERENCES

Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and Robustness of Penalized and Unpenalized Methods for Genetic Prediction of Complex Human Disease. *Genet Epidemiol*, **37**(2), 184–195.

Alexander, D. and Lange, K. (2011). Stability Selection for Genome-Wide Association. *Genet Epidemiol*, **35**, 722–728.

Barrett, J., Fry, B., Maller, J., and Daly, M. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Nat Rev Genet*, **21**, 263–265.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, **19**, 1212–1242.

Cantor, R., Lange, K., and Sinsheimer, J. (2010). Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *Am J Hum Genet*, **86**(1), 6–22.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and R-software hdi. *Statistical Science*, **30**, 533–558.

Franke, A. *et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*, **42**(12), 11181125.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1).

He, Q. and Lin, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics*, **27**(1), 1–8.

Hill, W. J. and Robertson, A. (1968). Linkage Disequilibrium in Finite Populations. *Theoretical and Applied Genetics*, **38**, 226–231.

Hoerl, A. and Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55–67.

Hoggart, C., Whittaker, J., Iorio, M. D., and Balding, D. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLOS Genetics*, **4**(7).

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice Hall, Upper Saddle River, NJ, USA.

Julia, A. *et al.* (2013). A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at RBX1-EP300. *Gut*, **62**(10), 1440–1445.

Kenny, E. *et al.* (2012). A Genome-Wide Scan of Ashkenazi Jewish Crohns Disease Suggests Novel Susceptibility Loci. *PLOS Genetics*, **8**(3).

Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics*, **27**(4), 516–523.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C., Davidson, R., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat Methods*, **8**, 833–835.

Malo, N., Libiger, O., and Schork, N. (2008). Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet*, **82**(2), 375–385.

Mandozzi, J. and Bühlmann, P. (2015). Hierarchical testing in the high-dimensional setting with correlated variables. *J Am Statist Assoc (published online DOI: 10.1080/01621459.2015.1007209)*.

Manolio, T. *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J., and Hirschhorn, J. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev Genet*, **9**, 356–369.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall, London.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, **95**, 265–278.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *J R Stat Soc Series B*, **72**, 417–473.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-Values for High-Dimensional Regression. *JASA*, **104**(488), 1671–1681.

Panagiotou, O. and Ioannidis, J. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol*, **41**(1), 273–286.

Parkes, M. *et al.* (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohns disease susceptibility. *Nat Genet*, **39**(7), 830–832.

Plagnol, V. *et al.* (2011). Genome-Wide Association Analysis of Autoantibody Positivity in Type 1 Diabetes Cases. *PLoS Genet*, **7**(8).

Purcell, S. *et al.* (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Int J Epidemiol*, **81**(3), 559575.

Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K. (2013). A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, **29**(2), 206–214.

Rioux, J. *et al.* (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet*, **39**(5), 596–604.

Schork, N. (2001). Genome partitioning and whole-genome analysis. *Adv Genet*, **42**, 299–322.

Scott, L. *et al.* (2007). A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science*, **316**(5829), 1341–1345.

Shi, G., Boerwinkle, E., Morrison, A., Gu, C., Chakravarti, A., and Rao, D. (2011). Mining Gold Dust Under the Genome Wide Significance Level: A Two-Stage Approach to Analysis of GWAS. *Genet Epidemiol*, **35**(2), 111–118.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661–678.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*, **58**, 267–288.

Welter, D. *et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, **42**.

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genet Epidemiol*, **34**(3), 275–285.

Yang, J. *et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, **42**, 565–569.

Yang, J., Lee, S., Goddard, M., and Visscher, P. (2011). GCTA: a tool for Genome-wide Complex Trait Analysis. *Am J Human Genet*, **88**, 76–82.

Yang, J., Zaitlen, N., Goddard, M., Visscher, P., and Price, A. (2014). Mixed model association methods: advantages and pitfalls. *Nat Genet*, **46**, 100–106.

Zeggini, E. *et al.* (2007). Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. *Science*, **316**(5829), 1336–1341.

Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J R Stat Soc Series B Stat Methodol*, **76**, 217–242.