BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 18 2011, pages 2578–2585
doi:10.1093/bioinformatics/btr424

Genetics and population analysis

Advance Access publication July 22, 2011

Identifying disease-associated SNP clusters via contiguous outlier detection

Can Yang[1], Xiaowei Zhou[1], Xiang Wan[1], Qiang Yang[2], Hong Xue[3] and Weichuan Yu[1,*]

[1]Laboratory for Bioinformatics and Computational Biology, Department of Electronic and Computer Engineering, [2]Department of Computer Science and Engineering and [3]Division of Life Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Although genome-wide association studies (GWAS) have identified many disease-susceptibility single-nucleotide polymorphisms (SNPs), these findings can only explain a small portion of genetic contributions to complex diseases, which is known as the missing heritability. A possible explanation is that genetic variants with small effects have not been detected. The chance is <8% that a causal SNP will be directly genotyped. The effects of its neighboring SNPs may be too weak to be detected due to the effect decay caused by imperfect linkage disequilibrium. Moreover, it is still challenging to detect a causal SNP with a small effect even if it has been directly genotyped.

Results: In order to increase the statistical power when detecting disease-associated SNPs with relatively small effects, we propose a method using neighborhood information. Since the disease-associated SNPs account for only a small fraction of the entire SNP set, we formulate this problem as Contiguous Outlier DEtection (CODE), which is a discrete optimization problem. In our formulation, we cast the disease-associated SNPs as outliers and further impose a spatial continuity constraint for outlier detection. We show that this optimization can be solved exactly using graph cuts. We also employ the stability selection strategy to control the false positive results caused by imperfect parameter tuning. We demonstrate its advantage in simulations and real experiments. In particular, the newly identified SNP clusters are replicable in two independent datasets.

Availability: The software is available at: http://bioinformatics.ust.hk/CODE.zip.

Contact: eeyu@ust.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 9, 2011; revised on June 27, 2011; accepted on July 12, 2011

1 INTRODUCTION

For the purpose of understanding complex diseases, genome-wide association studies (GWAS) use high-throughput technologies to assay hundreds of thousands of single-nucleotide polymorphisms (SNPs). By October 2010, about 700 GWAS covering 150 diseases and traits have been published and nearly 3000 SNPs have been reported as significantly associated with diseases or traits (Baker, 2010).

Despite of the success of GWAS, most of the findings only explain a small portion of genetic contributions to complex diseases. For example, 18 SNPs identified from type 2 diabetes (T2D) only account for 6% of the inherited genetic risk. This phenomenon is referred to as the missing heritability (Manolio *et al.*, 2009). To find the missing heritability, some possible methods have been suggested from biological views (Eichler *et al.*, 2010; Manolio *et al.*, 2009), such as detecting genetic variants with small or moderate effects, identifying rare genetic variants and structural variants that contribute to the disease risk and detecting gene–gene interactions underlying the human diseases (Cordell, 2009). To speed up the search of genetic variants that underly human diseases, more powerful computational and statistical tools are needed.

It is typical to test about 500*K* SNPs simultaneously in GWAS. For large-scale inference, controlling the false discovery rate (FDR) turns out to be a powerful strategy (Benjamini and Hochberg, 1995; Efron, 2010). The most commonly used FDR procedure is called BH procedure (Benjamini and Hochberg, 1995), which is based on thresholding the ranked *P*-values. It has been observed that BH procedure is able to control the FDR at nominal level for case–control studies, but its power decreases as the dependency among SNPs increases (Sabatti *et al.*, 2003).

A better FDR control can be achieved in SNP analysis by exploring the dependency structure (e.g. linkage disequilibrium). A direct simulation approach (DSA) based on multivariate normal distribution (MVN) was proposed in Conneely and Boehnke (2007); Seaman and Muller-Myhsok (2005). Although these studies primarily focused on datasets used in candidate gene studies, their work suggested that the block-wise strategy might be a possible approach for genome-wide studies. By extending the framework of MVN, a sliding-window approach was further proposed to account for the dependency among locally intercorrelated markers (Han *et al.*, 2009). To be more adaptive to the dependency structure of the human genome, a hidden Markov chain Model (HMM), which is a more flexible structure compared with the block-wise and sliding-window structures, was proposed to account for the dependence (Wei *et al.*, 2009). However, applying these methods to analyze real SNP datasets in GWAS is very time consuming. Moreover, some methods report many false positive results in real applications. Thus, finding the missing heritability is still a challenging issue.

[*]To whom correspondence should be addressed.

In this article, we propose a computational method to detect disease-associated SNPs with relatively small effects:

(1) The chance is less than 8% that a causal SNP is directly genotyped. The effect sizes of their neighborhood SNPs may become weak due to their imperfect linkage disequilibrium with the causal one (Spencer *et al.*, 2011).

(2) Except for a few risk SNPs with relatively large effects, most of the disease-associated SNPs show small effects (odds ratio< 1.5) (Altshuler *et al.*, 2008; Hindorff *et al.*, 2009; Manolio, 2010). Due to limited sample size, it is still statistically challenging to detect SNPs with small effect sizes even though they are directly observed.

We assume that the disease-associated SNPs account for only a small portion of the entire SNP set (i.e. the sparsity assumption). Therefore, they can be considered as outliers. On one hand, the risk SNPs with small effects can only be detected as a group due to the limited sample size. On the other hand, adjacent SNPs tend to form a local SNP cluster due to the linkage disequilibrium (LD) structure of human genome. If there exists a causal SNP which is not directly observed, the test statistics of its neighboring SNPs would tend to be different from zero. These facts motivate us to impose a spatial continuity constraint for outlier detection. Hence, we formulate the problem as Contiguous Outlier DEtection (CODE) and exactly solve it by graph cuts. As the solution depends on two parameters for controlling sparsity and spatial continuity, we present our heuristics for these parameter tuning. We further use stability selection to reduce the effects of imperfect parameter tuning.

The rest of this article is organized as follows. In Section 2, we present our formulation, algorithm and other issues such as parameter tuning and stability selection. Section 3 discusses some closely related methods. Section 4 reports experiment results. Finally, we conclude our article in Section 5.

2 METHOD

2.1 Problem statement

Suppose we have a dataset with $\mathcal{L}$ SNPs in a case–control study. The z value of each SNP can be easily obtained using the Cochran-Armitage trend test or univariate logistic regression (see the Supplementary Material). Let $\mathbf{z}=\{z_1,z_2\dots,z_{\mathcal{L}}\}$ be a set of z values. For disease-unassociated SNPs, their z-values asymptomatically follow the standard normal distribution given a finite sample size. For disease-associated SNPs, their z-values will be more significantly different from zero as the sample size increases (Wang *et al.*, 2005). Based on the set of z values, the SNPs can be partitioned into two groups (Efron, 2008):

- The null group $\mathcal{G}_0$: SNPs are unassociated with the disease.
- The non-null group $\mathcal{G}_1$: SNPs are associated with the disease.

Our goal is to classify $\mathcal{L}$ SNPs into these two groups.

Throughout this article, we will use the following norms of a vector $\mathbf{x}$: $\|\mathbf{x}\|_0$ denotes the ℓ_0-norm, which counts the number of non-zero entries. $\|\mathbf{x}\|_1=\sum_i|x_i|$ denotes the ℓ_1-norm. $\|\mathbf{x}\|_2=\sqrt{\sum_i x_i^2}$ and $\|\mathbf{x}\|_2^2=\sum_i x_i^2$ denote the ℓ_2-norm and the squared ℓ_2-norm, respectively.

2.2 Formulation

We make the following assumption: most SNPs come from the null group and the associated SNPs are considered as outliers of the entire SNP set. Thus, we formulate the above problem as an outlier detection problem using the mean-shift model (Maronna *et al.*, 2006):

$$\mathbf{z}=\boldsymbol{\gamma}+\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\gamma}=\{\gamma_1,\gamma_2,\dots,\gamma_{\mathcal{L}}\}$ and $\boldsymbol{\epsilon}=\{\epsilon_1,\epsilon_2,\dots,\epsilon_{\mathcal{L}}\}$. This model decomposes $\mathbf{z}$ into two parts: $\boldsymbol{\gamma},\boldsymbol{\epsilon}$. The non-zero entries of $\boldsymbol{\gamma}$ indicate associated SNPs and $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$.

Since most SNPs come from the null group, $\boldsymbol{\gamma}$ will be a sparse vector. This makes the decomposition in Equation (1) a well-posed problem. Correspondingly, we propose to solve the following minimization problem:

$$\min_{\boldsymbol{\gamma}}\ \frac{1}{2}\|\mathbf{z}-\boldsymbol{\gamma}\|_2^2+\lambda_1\|\boldsymbol{\gamma}\|_0. \tag{2}$$

Since we are interested in the associated SNPs, i.e. the non-zero entries of $\boldsymbol{\gamma}$, we introduce $\mathbf{s}=(s_1,\dots,s_{\mathcal{L}}), s_i\in\{0,1\}, i=1,\dots,\mathcal{L}$ as the support of $\boldsymbol{\gamma}$:

$$s_i=\begin{cases}0, & \text{if } \gamma_i=0\\ 1, & \text{if } \gamma_i\neq 0.\end{cases} \tag{3}$$

Suppose SNP A is a causal SNP of the disease, its z value z_A will be significantly different from zero. As adjacent SNPs tend to form a local SNP cluster due to the block-wise structure of the human genome, the z-values of its neighboring SNPs also tend to be different from zero, but the magnitudes of these z values will be smaller than z_A due to their imperfect LD with SNP A. In order to make use of the neighborhood information to increase the statistical power, it is necessary to introduce a spatial continuity constraint on $\mathbf{s}$. To do so, we firstly rewrite Equation (2) as a minimization problem over $\mathbf{s}$.

It is easy to see that, as long as $\gamma_i\neq 0$, we must have $\gamma_i=z_i$ to minimize Equation (2). Thus, Equation (2) has the same minimizer as the following energy function:

$$\min_{\boldsymbol{\gamma}}\ \frac{1}{2}\sum_{i:\gamma_i=0} z_i^2+\lambda_1\|\boldsymbol{\gamma}\|_0. \tag{4}$$

Let $\mathcal{P}_{\mathbf{s}}(\mathbf{z})$ be the orthogonal projection of $\mathbf{z}$ onto the support $\mathbf{s}$,

$$\mathcal{P}_{\mathbf{s}}(\mathbf{z})(i)=\begin{cases}0, & \text{if } s_i=0\\ z_i, & \text{if } s_i=1\end{cases} \tag{5}$$

and $\mathcal{P}_{\mathbf{s}^\perp}(\mathbf{z})$ be its complementary projection, i.e. $\mathcal{P}_{\mathbf{s}}(\mathbf{z})+\mathcal{P}_{\mathbf{s}^\perp}(\mathbf{z})=\mathbf{z}$. The first term in Equation (4) can be written in a compact form as $\frac{1}{2}\|\mathcal{P}_{\mathbf{s}^\perp}(\mathbf{z})\|_2^2$. Noticing that $\|\boldsymbol{\gamma}\|_0=\|\mathbf{s}\|_0$, we can rewrite Equation (4) as:

$$\begin{aligned}&\min_{\mathbf{s}}\frac{1}{2}\|\mathcal{P}_{\mathbf{s}^\perp}(\mathbf{z})\|_2^2+\lambda_1\|\mathbf{s}\|_0.\\ &\text{s.t. } s_i\in\{0,1\}, i=1,\dots,\mathcal{L}.\end{aligned} \tag{6}$$

Notice that $s_i=1$ indicates the i-th SNP is detected as an associated one. Now we introduce a spatial continuity regularizer on $\mathbf{s}$ to model the LD effect of neighboring SNPs. We propose to solve the following optimization problem:

$$\begin{aligned}&\min_{\mathbf{s}}\frac{1}{2}\|\mathcal{P}_{\mathbf{s}^\perp}(\mathbf{z})\|_2^2+\lambda_1\|\mathbf{s}\|_0+\lambda_2\sum_{i=1}^{\mathcal{L}-1}w_i|s_i-s_{i+1}|.\\ &\text{s.t. } s_i\in\{0,1\}, i=1,\dots,\mathcal{L}.\end{aligned} \tag{7}$$

The fused term $\lambda_2\sum_{i=1}^{\mathcal{L}-1}w_i|s_i-s_{i+1}|$ encourages the associated SNPs to be detected in a block-wise manner by penalizing the first-order difference. Our assumption is that the first-order model is a reasonably good approximation of an LD block. This is similar to the ideas of HMM and the fused Lasso (Tibshirani and Wang, 2008). We allow different weights between adjacent SNPs to accommodate different local LD effects. Specifically, let $r_{i,i+1}$ be the correlation of SNP i and SNP $i+1$. Here we use $r_{i,i+1}^2$ as w_i. In fact, $r_{i,i+1}^2$ is the composite LD value of SNP i and SNP $i+1$, which is a robust surrogate of the standard LD measure r^2 when the linkage phase is unknown (Schaid, 2004).

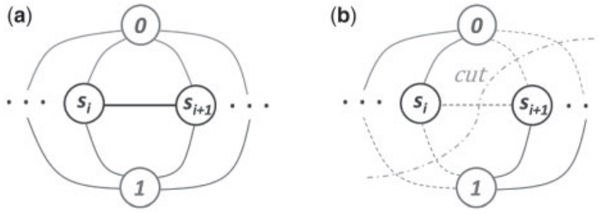

Fig. 1. (**a**) The constructed graph. (**b**) An example cut.

2.3 Algorithm

Now we investigate how to minimize the energy in Equation (7) over **s**. Noticing that $s_i \in \{0,1\}$, the energy can be rewritten as follows:

$$\frac{1}{2}\|\mathcal{P}_{\mathbf{s}^{\perp}}(\mathbf{z})\|_2^2+\lambda_1\|\mathbf{s}\|_0+\lambda_2\sum_{i=1}^{\mathcal{L}-1}w_i|s_i-s_{i+1}|$$

$$=\frac{1}{2}\sum_i\left[z_i^2(1-s_i)\right]+\lambda_1\sum_i s_i+\lambda_2\sum_{i=1}^{\mathcal{L}-1}w_i|s_i-s_{i+1}|. \qquad (8)$$

The above energy is in the standard form of the first-order Markov random fields (MRFs) with binary labels, which can be solved exactly in polynomial time using graph cuts (Boykov *et al.*, 2001; Kolmogorov and Zabih, 2004).

Next, we will describe the energy minimization via graph cuts briefly. The idea is to represent the energy by a graph and minimize the energy by finding the minimum cut of the graph. An example of an undirected graph $\mathcal{G}$ is as shown in Figure 1a. Graph nodes correspond to variables $s_i \in \mathbf{s}$, and two terminal nodes are added indicating binary labels: a 0-terminal and a 1-terminal. There are two types of undirected edges in $\mathcal{G}$: the neighboring link that connects each pair of nodes corresponding to neighboring variables in **s** and the terminal link that connects each variable node to two terminal nodes. In order to construct a graph representing Equation (8), we assign a weight to each link as follows:

$$\begin{cases} e_{\{s_i,s_{i+1}\}} \leftarrow \lambda_2 w_i, & i=1,\ldots,\mathcal{L}-1 \\ e_{\{s_i,0\}} \leftarrow \lambda_1, & i=1,\ldots,\mathcal{L} \\ e_{\{s_i,1\}} \leftarrow \frac{1}{2}z_i^2, & i=1,\ldots,\mathcal{L} \end{cases}, \qquad (9)$$

where $e_{\{s_i,s_{i+1}\}}$ is the weight of the link between s_i and s_{i+1}, $e_{\{s_i,0\}}$ is the weight of the link between s_i and the 0-terminal and $e_{\{s_i,1\}}$ is the weight of the link between s_i and the 1-terminal.

A cut of $\mathcal{G}$ is to partition $\mathcal{G}$ into two subgraphs by 'cutting' some edges such that two terminals are separated completely. The cost of a cut is defined as the sum of the weights of the cut edges. Figure 1b shows an example of a graph cut, in which $s_1,\ldots,s_i$ are connected to the 0-terminal and $s_{i+1},\ldots,s_{\mathcal{L}}$ are connected to the 1-terminal after the cut. When the weight is assigned as Equation (9), the cost of this cut is $\left(\sum_{j=1}^{j=i}e_{\{s_j,1\}}\right)+\left(\sum_{j=i+1}^{j=\mathcal{L}}e_{\{s_j,0\}}\right)+e_{\{s_i,s_{i+1}\}}$, which exactly equals to the energy value in Equation (8) for $s_j=0, j=1,\ldots,i$ and $s_j=1, j=i+1,\ldots,\mathcal{L}$.

Generally, two facts can be proved (Boykov *et al.*, 2001; Kolmogorov and Zabih, 2004): (i) there is a one-to-one mapping between a cut and a configuration of **s**, since any s_i will be connected to either the 0-terminal or the 1-terminal after the cut. (2) For the corresponding cut and configuration of **s**, the cut cost equals to the energy value in Equation (8). Note that the cut cost comes from three sources: if s_i is connected to the 0-terminal, $e_{\{s_i,1\}}$ will be cut and $\frac{1}{2}z_i^2$ is added to the cost. If s_i is connected to the 1-terminal, $e_{\{s_i,0\}}$ will be cut and λ_1 is added to the cost. If neighboring s_i and s_{i+1} are connected to different terminals, $e_{\{s_i,s_{i+1}\}}$ will be cut and $\lambda_2 w_i$ is added to the cost. The sum of these costs equals the energy in Equation (8). Thus, the minimizer of Equation (8) $\mathbf{s}^*$ can be obtained by finding the cut with the minimal cost, which can be solved efficiently using the standard max-flow algorithm (Cormen, 2001).

2.4 Parameter tuning

We need to specify two parameters λ_1 and λ_2 in our method, where λ_1 controls the sparsity of **s** and λ_2 controls its spatial continuity. Correctly choosing λ_1 and λ_2 can reduce false positive result and also increase statistical power. When $z_1,\ldots,z_{\mathcal{L}}$ are independent and identically Gaussian distributed variables, then $z_i \sim \mathcal{N}(0,\sigma^2), i=1,\ldots,\mathcal{L}$, a simple choice for λ_1 is $\lambda_1=\sigma\sqrt{2\log(\mathcal{L})}$ which is derived in Donoho and Johnstone (1994). This choice is due to the fact that the expected maximum of $|z_i|, i=1,\ldots,\mathcal{L}$ is approximately $\sigma\sqrt{2\log(\mathcal{L})}$ when $z_i \sim \mathcal{N}(0,\sigma^2), i=1,\ldots,\mathcal{L}$.

The above formula assumes that there are $\mathcal{L}$ independent SNPs. In reality, the effective number of SNPs is smaller than $\mathcal{L}$ due to the LD effect. Recently, Han *et al.* (2009) used a permutation-based method to estimate the effective number of SNPs. According to their estimation, the effective number of the 2.7 million HapMap SNPs is about one million. Thus, We set $\lambda_1=\sigma\sqrt{2\log(\mathcal{L}/\mathcal{E})}$, where $\mathcal{L}/\mathcal{E}$ is the effective number of SNPs with $\mathcal{E}=2.7$. [1]

Next, we empirically choose λ_2 in the following way: λ_2 begins with a large value $\lambda_2^{\max}$. Then we solve problem (7) using graph cuts and obtain the optimal solution of **s**. After that, we get the residual $\mathbf{r}=\mathcal{P}_{\mathbf{s}^{\perp}}(\mathbf{z})$ and compute the variance of **r**, denoted as var(**r**). We gradually decrease λ_2 such that var(**r**) gets close to σ^2. Specifically, when var$(\mathbf{r})>\sigma^2$, we decrease λ_2 by a factor η: $\lambda_2 \leftarrow \eta\lambda_2$. We repeat this process until var$(\mathbf{r})-\sigma^2<\varepsilon$. The idea behind this procedure is that the variance of $z_i, i \in \mathcal{G}_0$ (z values of the null group) should be close to σ^2. In our experiment, we set $\lambda^{\max}=600$ ($\lambda_2^{\max}$ can be easily chosen as long as most of **s** are zero. $\lambda_2^{\max}=600$ is satisfactory in general), $\eta=0.9$ and $\varepsilon=0.01$.

The remaining issue is about the parameter σ. Here we would like to estimate σ using empirical Bayesian inference as proposed in Efron (2010). We assume $z_i \sim \mathcal{N}(\delta,\sigma^2), \forall i \in \mathcal{G}_0$. The choice of $\delta=0$ and $\sigma=1$ implies that the theoretical null distribution is used in statistical inference. However, the value of σ can be quite different in large-scale inference, as pointed out by Efron (2004). Instead of directly using the theoretical null distribution, Efron (2004) proposed to estimate the empirical null distribution. Regarding to SNP data analysis, we employ the *locfdr* algorithm (Efron, 2010) to estimate δ and σ, and remove the mean of **z**: $z_i \leftarrow z_i-\delta, \forall i$. In practice, we find that the estimated δ and σ are very close to 0 and 1, respectively. This means that the empirical null distribution is very close to the theoretical null distribution. Our observation is consistent with the result in Efron (2010) when the *locfdr* algorithm was applied to SNP data analysis.

2.5 Stability selection

Although the heuristic for parameter tuning sounds reasonable, it may produce too many false positive results. To reduce false positives, we employ the stability selection strategy (Meinshausen and Buhlmann, 2010) to reduce the effect of parameter tuning. Using stability selection, detection of a SNP as an associated SNP does not rely on a single run of a particular parameter setting, but depends on the probability that it is detected under model perturbation via subsampling.

Specifically, for each subsamling round, we randomly sample half of the cases and half of the controls from the entire dataset. Half subsampling is very close to the Bootstrap method, which has been theoretically analyzed in Buhlmann and Yu (2002); Friedman and Hall (2007). Let $\mathbf{z}_b^*$ denote the set of z values for the b-th subsampling. We can efficiently obtain $\mathbf{z}_b^*$ using the Cochran–Armitage trend test that can be implemented using Boolean operations (Wan *et al.*, 2010). After that, we run our model on $\mathbf{z}_b^*$ and obtain the support $\mathbf{s}_b^*$. Note that $s_{i,b}^*=1$ indicates that the i-th SNP is detected in the b-th subsampling. Therefore, the probability of the i-th SNP being detected

[1] Notice that the change of $\mathcal{E}$ only makes λ_1 slightly different, e.g. when $\mathcal{L}=30000$ and $\sigma=1$, we have $\lambda_1 \simeq 4.32, 4.29$ and 4.22 for $\mathcal{E}=2.7, 3$ and 4, respectively. In this sense, the increasing number of reported SNPs and the correspondingly change of $\mathcal{E}$ do not make λ_1 too much difference.

can be easily obtained by

$$\pi_i=\frac{\sum_{b=1}^{B} s_{i,b}^{*}}{B}, \tag{10}$$

where B is the number of subsampling. Typically, we set $B=100$ as in Meinshausen and Buhlmann (2010). We can obtain a set of interesting SNPs as $\mathcal{A}_\tau=\{i:\pi_i\geq\tau\}$, where τ is a threshold. The number of SNPs in $\mathcal{A}_\tau$ is denoted as $|\mathcal{A}_\tau|$.

2.6 Estimate of FDR

It has been shown that the false positive error can be controlled using this strategy (Meinshausen and Buhlmann, 2010) for the Lasso model (Tibshirani, 1996). However, this theoretical result cannot be directly applied here due to our different formulation (7). Thus, we need to estimate the FDR of $\mathcal{A}_\tau$ for a given threshold τ.

Although we do not have independent hypothesis for each SNP, we can still use

$$\widehat{\mathrm{FDR}}_\tau=\frac{\mathcal{N}_\tau}{|\mathcal{A}_\tau|} \tag{11}$$

as a rough estimator for FDR (Efron *et al.*, 2001; Storey and Tibshirani, 2003; Tibshirani and Wang, 2008), where $\mathcal{N}_\tau$ is the number of SNPs picked at threshold τ under the null distribution. We can use permutation to obtain the number of SNPs picked under the null distribution. Specifically, for a given threshold τ, we do T permutations. During the t-th permutation, we permute the case–control label to generate a null dataset, denoted as $\mathcal{D}^{(t)}$. Then, we run B times subsampling on $\mathcal{D}^{(t)}$. For each subsampling, we use the heuristic for parameter tuning and solve problem (6). After B times subsampling, we obtain $\boldsymbol{\pi}^{(t)}$ according to Equation (10). For the given threshold τ, we have $\widetilde{\mathcal{A}}_\tau^{(t)}=\{i:\pi_i^{(t)}\geq\tau\}$. Then the final estimation of $\widehat{FDR}_\tau$ is given by

$$\widehat{FDR}_\tau=\frac{\mathcal{N}_\tau}{|\mathcal{A}_\tau|}=\frac{\frac{1}{T}\sum_{t=1}^{T}|\widetilde{\mathcal{A}}_\tau^{(t)}|}{|\mathcal{A}_\tau|}, \tag{12}$$

where $|\widetilde{\mathcal{A}}_\tau^{(t)}|$ denotes the number of SNPs in $\widetilde{\mathcal{A}}_\tau^{(t)}$.

2.7 Analysis of multiple chromosomes

Chromosomes are inherited independently based on Mendel's Law for most diseases. In this article, we treat different chromosomes separately. Suppose there are K chromosomes. Let $\mathcal{L}_k$ be the number of SNPs of the k-th chromosome and $\mathbf{z}^{(k)}=\{z_1^{(k)},\ldots,z_{\mathcal{L}_k}^{(k)}\}$ be their z values. We have $\sum_{k=1}^{K}\mathcal{L}_k=\mathcal{L}$. There will be no fused terms between different chromosomes. Accordingly, we solve the following K optimization problems separately:

$$\min_{\mathbf{s}^{(k)}}\frac{1}{2}\|\mathcal{P}_{\mathbf{s}^{(k)\perp}}(\mathbf{z}^{(k)})\|_2^2+\lambda_1\|\mathbf{s}^{(k)}\|_0+\lambda_2\sum_{i=1}^{\mathcal{L}_k-1}w_i|s_i^{(k)}-s_{i+1}^{(k)}|. \tag{13}$$
$$\text{s.t. } s_i^{(k)}\in\{0,1\}, i=1,\ldots,\mathcal{L}_k.$$

For parameter tuning, we set $\lambda_1=\sigma_k\sqrt{\log(\mathcal{L}/2.7)}$, where σ_k is estimated from the k-th chromosome. We use $\mathcal{L}$ rather than $\mathcal{L}_k$ for setting λ_1 because we are analyzing multiple chromosomes. The strategy for tuning λ_2 is the same as what we described in Section 2.4.

3 RELATIONSHIP BETWEEN OUR METHOD AND OTHER METHODS

Fused Lasso (Tibshirani *et al.*, 2005) is a closely related method and it has been used in hot spot detection for CGH data (Tibshirani and Wang, 2008). To analyze the SNP data using the idea of the fused Lasso, the observed z values can be decomposed as:

$$\mathbf{z}=\boldsymbol{\mu}+\boldsymbol{\epsilon}. \tag{14}$$

The fused Lasso (Tibshirani *et al.*, 2005) could be modified slightly as follows:

$$\min_{\boldsymbol{\mu}}\sum_{i=1}^{\mathcal{L}}(z_i-\mu_i)^2+\lambda_1\sum_{i=1}^{\mathcal{L}}|\mu_i|+\lambda_2\sum_{i=1}^{\mathcal{L}-1}w_i|\mu_i-\mathrm{sign}(r_{i,i+1})\mu_{i+1}| \tag{15}$$

where $r_{i,i+1}$ is the correlation of SNP i and SNP $i+1$, sign$(r_{i,i+1})$ is used to adjust the sign difference between μ_i and μ_{i+1}, and w_i is the weight for the i-th fused term. Similarly, w_i can be used to accommodate the local correlation structure of adjacent SNPs, e.g. $w_i=r_{i,i+1}^2$. Notice that each element of $\boldsymbol{\mu}$ is a real number while each element of $\mathbf{s}$ in Equation (7) can only be 0 or 1. Interestingly, both the final solutions of Equation (15) $\hat{\boldsymbol{\mu}}$ and of Equation (7) $\hat{\mathbf{s}}$ are piecewise constant. As pointed out by Rinaldo (2009), the final solution of Equation (15) $\hat{\boldsymbol{\mu}}$ is a biased estimation due to the shrinkage effect of ℓ_1 regularization (Hastie *et al.*, 2009; Mazumder *et al.*, 2011).

Since $\mathbf{s}$ is discrete, our formulation (7) can be rewritten as ℓ_0 regularization:

$$\min_{\mathbf{s}}\frac{1}{2}\|\mathcal{P}_{\mathbf{s}^\perp}(\mathbf{z})\|_2^2+\lambda_1\|\mathbf{s}\|_0+\lambda_2\sum_{i=1}^{\mathcal{L}-1}w_i|s_i-s_{i+1}|_0. \tag{16}$$
$$\text{s.t. } s_i\in\{0,1\}, i=1,\ldots,\mathcal{L}.$$

where

$$|s_i-s_{i+1}|_0=\begin{cases}1, & \text{if } s_i\neq s_{i+1}\\ 0, & \text{if } s_i=s_{i+1}\end{cases}. \tag{17}$$

It is known that ℓ_0 regularization and ℓ_1 regularization have different mathematical properties (Mazumder *et al.*, 2011; She, 2009): compared with ℓ_1 regularization, ℓ_0 regularization gives unbiased estimation but larger variance. The performance of these two formulations depends on the problem at hand.

The *locfdr* method (Efron, 2010) extends the original FDR (Benjamini and Hochberg, 1995) to the local FDR using the two-group model (the relationship between local FDR and FDR is given in the Supplementary Material). It assumes that z_i with $i\in\mathcal{G}_0$ comes from the null distribution $f_0(z|\boldsymbol{\theta}_0)$ with probability p_0 and others come from the alternative distribution $f_1(z|\boldsymbol{\theta}_1)$ with probability $p_1=1-p_0$, where $\boldsymbol{\theta}_0,\boldsymbol{\theta}_1$ are parameters of the distributions f_0 and f_1. Under mild assumptions, $p_0,\boldsymbol{\theta}_0,p1,\boldsymbol{\theta}_1$ can be accurately estimated from data. It turns out that this simple two-group model works well in practice. The local index of significance (LIS) method (Wei *et al.*, 2009) can be considered as an extension of the *locfdr* method. Specifically, let $\mathcal{I}_i$ be the group indicator of the i-th SNP, $\mathcal{I}_i=0$ if $i\in\mathcal{G}_0$ and $\mathcal{I}_i=1$ otherwise. HMM is used to model the dependence of the adjacent indicators $\mathcal{I}_i$ and $\mathcal{I}_{i+1}$. It further assumes that z_i with $i\in\mathcal{G}_0$ comes from the null distribution $f_0(z|\boldsymbol{\theta}_0)$ and others are from the alternative distribution $f_1(z|\boldsymbol{\theta}_1)$. All the parameters are estimated using expectation maximization (EM) algorithm. Wei *et al.* (2009) has shown that this model is particularly powerful for identifying non-null SNP clusters. We will compare CODE with these methods in simulation studies.

4 RESULTS

4.1 Simulation studies

In this section, we compare our method with the *locfdr* method (Efron, 2004), LIS (Wei *et al.*, 2009) and the modified fused Lasso [Equation (15)]. In order to simulate more realistic linkage disequilibrium patterns of SNP data, we choose 3000 individuals from the WTCCC control dataset (Consortium, 2007) as the pool. We apply the standard quality control procedure (the proportion of miss values $\leq 10\%$, the minor allele frequency $\geq 5\%$ and the P-value of Hardy–Weinberg equilibrium ≥ 0.0001) to pre-process this dataset. We use SNPs from chromosome 10 for our simulation studies. We setup two disease causal regions of this chromosome in

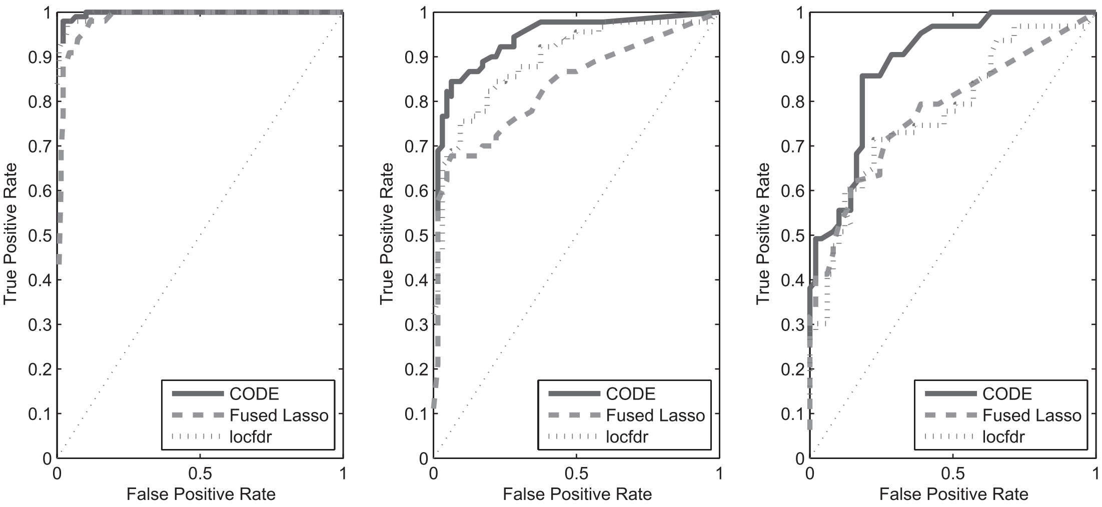

Fig. 2. Performance comparison of CODE, the modified fused Lasso and the *locfdr* method. From left to right, ORs are 1.49, 1.35 1.28, respectively.

our simulation. This two regions are chosen by the following criteria: first, within the region, the LD structure is well maintained. Second, the two regions are distantly separated such that their causal signals do not interfere with each other. After analyzing the LD pattern of the WTCCC data, we choose the 2000-th SNP rs1575568 and the 5000-th SNP rs2985526 to be the causal SNPs in our simulation. Then we generate case–control data using the following logistic regression model:

$$\Pr(Y=1|X_1,X_2)=\frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2)}, \tag{18}$$

where Y is the case–control label, X_1,X_2 are the genotype corresponding to ‘rs1575568’ and ‘rs2985526’, respectively. Here we have assumed the additive model of the disease risk on the log scale. In order to simulate small or moderate effects of the causal SNPs, we set $\beta_1=\beta_2=0.25, 0.3, 0.4$ corresponding to odds ratio (OR) 1.28, 1.35, 1.49, respectively. For fixed β_1,β_2, we adjust β_0 such that the simulated case–control data is balanced. In order to mimic the situation that the causal SNPs are not directly genotyped, we remove these two causal SNPs. To avoid the case that the causal SNPs have been well tagged, we further remove the SNPs which are almost perfectly correlated with them ($r^2\geq 0.95$).

When the causal variant is not included, the definition of true and false positives may be problematic. Here all the identified SNPs, which are close to and highly correlated with the causal one ($0.5\leq r^2\leq 0.95$), are recognized as true positives, otherwise as false positives. The performance of these methods are show in Figure 2. When the signal is moderate ($OR=1.49$), all the methods shows similar performance. As the signal gets weaker and weaker, CODE is clearly the winner followed by the *locfdr* method and the modified fused Lasso.

It is not surprising that CODE outperforms the *locfdr* method. On one hand, the spatial continuity constraint enables a SNP with the weak effect to be detected by borrowing the strength from its neighbors. On the other hand, this constraint prevents a strong noise signal from being detected if its neighborhood signals are all weak.

However, it seems surprising that the modified fused Lasso does not perform as well as CODE since it also uses neighborhood information. Our explanation is based on the following facts:

- The estimation obtained from ℓ_1 regularization is biased but with smaller variance.
- The estimation obtained from ℓ_0 regularization is unbiased but with larger variance.

For CODE, which uses ℓ_0 regularization, the large variance has been greatly reduced by the stability selection strategy, where subsampling and aggregation are used like ‘Bagging’ (Breiman, 1996). For the fused Lasso, the improvement of reducing variance is limited while bias remains.

There is a possibility that we may underestimate the power of the modified fused Lasso due to parameter tuning. The heuristics of parameter tuning is not suitable for the modified fused Lasso. Thus, we choose Bayesian Information Criterion (BIC) to select a good model for the modified fused Lasso. Here arises the issue of the degree of freedom (DF). Recall that the number of non-zero blocks in the solution $\hat{\boldsymbol{\mu}}$ is defined as DF of the original fused Lasso (Tibshirani *et al.*, 2005). Since a different weight w_i is used for each fused term in the modified fused Lasso, the number of non-zero blocks can only be an approximation of the true DF. This inexact DF may lead to imperfect model selection with BIC. As we have used the stability selection strategy, which makes the final result insensitive to parameter tuning, the underestimation of the power of the modified fused lasso should be small.

We do not show the result of LIS in the main article, because we find that LIS produces too many false positives (Supplementary Material). Probably, this is because the EM algorithm gets stuck in a local optimum when estimating the parameters in HMM.

4.2 Calibration of the threshold of the selection probability based on empirical studies

We have proposed a procedure to estimate FDR (Benjamini and Hochberg, 1995) of our method in Section 2.6. For a given threshold τ, T permutations are involved (e.g. $T=100$) to estimate $\widehat{\mathrm{FDR}}_\tau$,

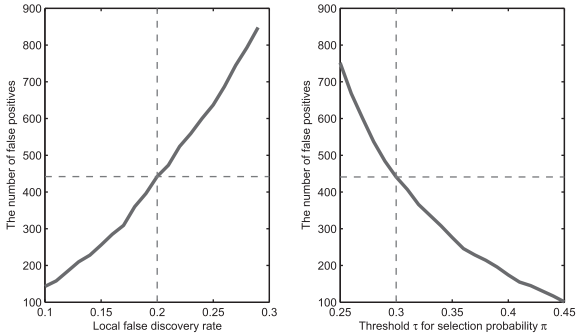

Fig. 3. Calibration of the threshold τ of the selection probability. We recommend $\tau = 0.30$ as the default setting as it corresponds to local false discovery rate 0.2 based on our empirical studies.

which is computationally intensive. A lower threshold τ may produce more false positives and a higher threshold τ corresponds to less false positives but may miss some true positives. Here we would like to conduct some empirical studies to connect the threshold τ with the local FDR of the *locfdr* method. In practice, this empirical calibration may provide a guideline for users to choose a suitable threshold τ.

Here we still use the WTCCC control dataset with 3000 individuals. Based on this real data, we run $T = 1000$ permutations. Specifically, at the t-th permutation, we randomly assign the case–control label for each individual to generate the t-th null dataset $\widetilde{\mathcal{D}}_t$, and then we obtain a set of z values of $\widetilde{\mathcal{D}}_t$, denoted as $\widetilde{\mathbf{z}}_t$. We run the $\widetilde{locfdr}$ method on $\widetilde{\mathbf{z}}_t$ and record the local FDR of $\widetilde{\mathbf{z}}_t$, denoted as $\widetilde{\mathbf{locfdr}}_t$. We also run our method on the null dataset $\widetilde{\mathcal{D}}_t$ by $B = 100$ subsampling, and then we obtain $\widetilde{\boldsymbol{\pi}}_t$ (Equation 10). After all T permutations, we obtain $\widetilde{\mathbf{locfdr}} = \{\widetilde{\mathbf{locfdr}}_1, \ldots, \widetilde{\mathbf{locfdr}}_T\}$ and $\widetilde{\boldsymbol{\pi}} = \{\widetilde{\boldsymbol{\pi}}_1, \ldots, \widetilde{\boldsymbol{\pi}}_T\}$. After that, we can apply different local FDRs to $\widetilde{\mathbf{locfdr}}$ and check how many false positives will be reported. Also, we can apply different threshold τ to $\widetilde{\boldsymbol{\pi}}$ and check the number of false positives.

Figure 3 provides the calibration result using the SNP data from chromosome 10. Under the local FDR 0.2, which is the default setting of the *locfdr* method, we have observed 442 false positives. For our method, 441 false positives have been observed when $\tau = 0.30$. The correspondence also holds for some other thresholds, for example, local false positive rate 0.15 (256 false positives reported by the *locfdr* method) roughly corresponds to $\tau = 0.36$ (246 false positives). We have observed that this relationship is rather stable when other SNP datasets are used. Thus, we recommend $\tau = 0.30$ as the default setting for our method.

4.3 Computational time

Computational efficiency is very important in real applications. We test the running time of our method on a desktop computer. The results are shown in Table 1.

The *locfdr* method has a computational advantage compared with CODE. It can finish the analysis of 30000 SNPs within 1 s after the z values are obtained. For the modified fused Lasso, we solve 100 combinations of $[\lambda_1, \lambda_2]$ using the split Bregman method (Ye and Xie, 2010) for each subsampling. It takes 2148 s for a dataset with $\mathcal{L} = 30,000, n = 5,000$. LIS takes about 563 s to finish the analysis.

Table 1. Computational time of CODE. The timing is carried out on one 3.0 GHz CPU with 4 GB memory running Windows XP professional system

Problem size	CPU time (s)
$\mathcal{L} = 30\,000, n = 5000$	45
$\mathcal{L} = 10\,000, n = 5000$	13

The reported CPU time is for $B = 100$ subsamplings. In each subsampling, it involves calculation of z values and solving multiple times of Equation (7) using graph cuts.

Thus, CODE has computational advantages over the fused Lasso and LIS.

4.4 Experiments on two independent datasets of Crohn's Disease

We apply our method to two independent dataset of Crohn's Disease (CD). One CD dataset comes from WTCCC (Consortium, 2007), in which about 500K SNPs are genotyped in about 5K subjects. We apply the standard quality control procedure to preprocess this dataset. The number of remaining SNPs is around 360 000. The other CD dataset comes from Duerr *et al.* (2006), in which 308 332 autosomal SNPs were assayed on the Illumina HumanHap300 chip. After the same quality control strategy, the number of remaining SNPs is 291 964. We call this dataset as 'Duerr's CD dataset'.

CODE and the *locfdr* method are applied to these two datasets. The results are listed in Tables 2 and 3. As we can see, these SNPs are identified by both methods and their corresponding genes are replicable in these two independent datasets. According to the data base search result from FunctSNP (Goodswen *et al.*, 2010), functional SNPs are indicated by '*' in the table. Genes *IL23R*, *NOD2*, *ATG16L1* and *ZNF365* have been confirmed to be Crohn's disease susceptibility genes (Consortium, 2007; Cummings *et al.*, 2007; Duerr *et al.*, 2006; Haritunians *et al.*, 2011; Wehkamp *et al.*, 2004). Although the analysis result does not provide new discoveries, it demonstrates that CODE and the *locfdr* method are very powerful because these replicable signals have not been detected in the original analysis (Duerr *et al.*, 2006). Regarding to Gene *CYLD* which is identified in these two datasets by both methods, our analysis result suggests that it might be a CD susceptibility gene although it has not been discussed in the literature.

Moveover, CODE has identified some new signals which are replicable in these two independent datasets. These results are given in Tables 4 and 5. For Duerr's CD dataset, the *locfdr* method suggests that rs1000141 is associated with CD with local FDR 0.1723. Our method identifies this SNP with selection probability 0.59 and SNP rs3792091 with selection probability 0.47. These two SNPs locate in gene *SAG*. For the WTCCC CD dataset, our method also identifies SNPs (rs2304773-rs2241873) which locate in gene *SAG* with selection probability 0.56 while the locfdr method consider them as insignificant ones. It is interesting that SNP rs2304773 is a functional SNP according to the data base search result of FunctSNP (Goodswen *et al.*, 2010). Based on this finding, we conjecture that Gene *SAG* is associated with CD. In addition, our method also identifies some other SNPs which locate genes *NKD1* and *PLD5* in both datasets. The biological interpretation of these findings remains unclear. It would be of great interest if their biological functions could be investigated.

Table 2. Disease-associated SNPs identified by CODE and the *locfdr* method from the WTCCC CD dataset and Duerr's CD dataset

SNP name	z	π_i	*locfdr*	Location	Gene	Chromosome
rs17375018	−5.8112	0.99	0.00713	67655147	IL23R	1p31.3
rs6664119	−4.6748	0.99	0.22985	67655895	IL23R	1p31.3
rs11805303	7.5464	0.99	4.00e-06	67675516	IL23R	1p31.3
rs10489629	−7.2616	0.99	5.00e-06	67688349	IL23R	1p31.3
rs2201841	7.3793	0.99	8.00e-06	67694202	IL23R	1p31.3
rs10210302	−7.6481	1.00	1.00e-06	234158839	ATG16L1	2q37.1
rs6752107	−7.6895	1.00	1.00e-06	234161448	ATG16L1	2q37.1
rs6431654	−7.8178	1.00	<1.00e-10	234161769	ATG16L1	2q37.1
rs3828309	−7.5160	1.00	4.00e-06	234180410	ATG16L1	2q37.1
rs3792106	−7.5155	1.00	4.00e-06	234190740	ATG16L1	2q37.1
rs10995271	5.1153	0.46	0.03923	64438486	ZNF365 (DS)	10q21.2
rs10761659	−4.8225	0.60	0.07791	64445564	ZNF365 (DS)	10q21.2
rs17221417	6.7597	1.00	0.00027	50739582	NOD2	16q21
rs17312836	−5.9375	1.00	0.00086	50741462	NOD2	16q21
rs2066843*	7.0801	1.00	4.00e-05	50745199	NOD2	16q21
rs1861759*	−5.7804	1.00	0.00212	50745583	NOD2	16q21
rs748855	−6.0248	1.00	0.00049	50751398	NOD2	16q21
rs1861758	−5.9306	1.00	0.00089	50751787	NOD2	16q21
rs2076756	7.7963	1.00	<1.00e-10	50756881	NOD2	16q21
rs3135499*	−6.3497	1.00	6.20e-05	50766127	NOD2	16q21
rs8060598	−6.5550	1.00	2.20e-05	50781802	CYLD	16q12.1
rs3785142	5.4603	1.00	0.08849	50787147	CYLD	16q12.1
rs7342715	6.2788	1.00	0.00354	50787483	CYLD	16q12.1
rs3135503	−6.4501	1.00	3.30e-05	50791250	CYLD	16q12.1
rs4785450	−6.0114	1.00	0.00052	50792268	CYLD	16q12.1

The genes located by these SNPs are replicable in these two data sets. The columns are the SNP name of an identified SNP, its z value given by the Cochran-Armitage trend test, the selection probability of our method, the local false discovery rate given by the *locfdr* method, the SNP position, the gene in which the SNP located, the chromosome position of the gene. We run $B=100$ subsampling in CODE to estimate the selection probability. The functional SNPs are indicated with $*$ in the column 'SNP name'. 'DS' in the column 'Gene' indicates that the gene is at the down stream of the given SNP.

Table 3. Disease-associated SNPs identified by CODE and the *locfdr* method from Duerr's CD dataset

SNP name	z	π_i	*locfdr*	Location	Gene	Chromosome
rs10489629	−6.0931	0.84	2.20e-05	67688349	IL23R	1p31.3
rs10889677*	6.2684	0.86	1.00e-05	67725120	IL23R	1p31.3
rs11209026*	−5.8267	0.99	0.00014	67705958	IL23R	1p31.3
rs11465804	−6.1212	0.97	1.90e-05	67702526	IL23R	1p31.3
rs1343151	−6.6332	0.98	<1.00e-10	67719129	IL23R	1p31.3
rs2201841	6.1547	0.94	1.50e-05	67694202	IL23R	1p31.3
rs3792106	−4.2932	0.47	0.19652	234190740	ATG16L1	2q37.1
rs2241880*	−5.3373	0.80	0.00711	234183368	ATG16L1	2q37.1
rs224136	−4.6647	0.51	0.13768	64470675	ZNF365 (DS)	10q21.2
rs10521209	−5.2431	0.82	0.00048	50755709	NOD2	16q21
rs11647841	−5.2195	0.80	0.00057	50743331	NOD2	16q21
rs8060598	−5.205	0.74	0.00063	50781802	CYLD	16q12.1

Table 4. Disease-associated SNPs identified from the WTCCC CD dataset

SNP name	z	π_i	*locfdr*	Location	Gene	Chromosome
rs695021	−1.1138	0.34	0.97522	242328222	PLD5	1q43
rs316896	3.8128	0.34	0.75375	242329814	PLD5	1q43
rs316895	3.6977	0.34	0.78432	242331083	PLD5	1q43
rs12139740	3.6126	0.34	0.80410	242332976	PLD5	1q43
rs316871	−1.2288	0.34	0.96938	242342884	PLD5	1q43
rs10926640	2.9637	0.34	0.90147	242343502	PLD5	1q43
rs12135329	3.3761	0.34	0.84681	242347587	PLD5	1q43
rs316823	−0.94208	0.34	1.00000	242356028	PLD5	1q43
rs10449290	3.0628	0.34	0.88963	242357338	PLD5	1q43
rs2304773*	−0.40343	0.56	1.00000	234235820	SAG	2q37.1
rs894100	−2.8705	0.56	1.00000	234237765	SAG	2q37.1
rs3792097	−2.5395	0.56	1.00000	234238784	SAG	2q37.1
rs3792096	−2.7996	0.56	1.00000	234238900	SAG	2q37.1
rs2241874	−3.7749	0.56	1.00000	234247627	SAG	2q37.1
rs2241873	−3.4969	0.56	1.00000	234247924	SAG	2q37.1
rs4785433	−1.3691	0.44	1.00000	50586941	NKD1	16q12
rs933566	2.2137	1.00	1.00000	50642201	NKD1	16q12
rs8047222	4.4660	1.00	0.56836	50660962	NKD1	16q12

Table 5. Disease-associated SNPs identified from Duerr's CD datasets

SNP name	z	π_i	*locfdr*	Location	Gene	Chromosome
rs6429357	−3.2291	0.36	0.56599	242633259	PLD5	1q43
rs6429332	3.5870	0.38	0.40225	242249484	PLD5 (US)	1q43
rs1000141	−4.3470	0.59	0.17236	234242347	SAG	2q37.1
rs3792091	−2.9304	0.47	0.77164	234251322	SAG	2q37.1
rs4785220	3.3683	0.34	0.45255	50627378	NKD1	16q12
rs4785437	−3.8216	0.35	0.33123	50598798	NKD1	16q12

'US' in the column 'Gene' indicates that the gene is at the upstream of the given SNP.

5 CONCLUSION

In order to detect disease-associated SNP clusters, our formulation makes use of information among adjacent SNPs. It turns out that our formulation is an ℓ_0 regularization problem and it can be exactly solved by graph cuts. We have shown that this method is particularly powerful when the effective size is small or moderate, which may be of great help in finding the missing heritability. Using two independent CD datasets, we demonstrate that CODE is able to reliably detect weak signals.

CODE begins with z values assumed to be normally distributed. When applying CODE to statistics with other distributions, they need to be transformed to be normally distributed as what has been done for the *locfdr* method (Efron, 2004).

ACKNOWLEDGEMENTS

We thank the two anonymous reviewers for their constructive comments, which greatly help us improve our manuscript.

Funding: Hong Kong University of Science and Technology (Grants RPC06/07.EG09, RPC07/08.EG25, RPC10EG04); Sir Michael and Lady Kadoorie Funded Research into Cancer Genetics.

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al.* (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.

Baker,M. (2010) Genomics: the search for association. *Nature*, **467**, 1135–1138.

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

Boykov,Y. *et al.* (2001) Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**, 1222–1239.

Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Buhlmann,P. and Yu,B. (2002) Analyzing bagging. *Ann. Stat.*, **30**, 927–961.

Conneely,K. and Boehnke,M. (2007) So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am. J. Hum. Genet.*, **81**, 1158–1168.

Consortium,T.W.T.C.C. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Cordell,H. (2009) Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Cormen,T. (2001) *Introduction to Algorithms*. The MIT press, Cambridge, MA.

Cummings,J. *et al.* (2007) Confirmation of the role of ATG16L1 as a Crohn's disease susceptibility gene. *Inflamm. Bowel Dis.*, **13**, 941–946.

Donoho,D. and Johnstone,J. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Duerr,R. *et al.* (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**, 1461–1463.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Am. Stat. Assoc.*, **99**, 96–104.

Efron,B. (2008) Microarrays, empirical Bayes and the two-groups model. *Stat. Sci.*, **23**, 1–22.

Efron,B. (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Eichler,E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Friedman,J. and Hall,P. (2007) On bagging and nonlinear estimation. *J. Stat. Plann. Inf.*, **137**, 669–683.

Goodswen,S. *et al.* (2010) FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases. *BMC Bioinformatics*, **11**, 311.

Han,B. *et al.* (2009) Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.*, **5**, e1000456.

Haritunians,T. *et al.* (2011) Variants in ZNF365 isoform D are associated with Crohn's disease. *Gut.*, **60**, 1060–1067.

Hastie,T. *et al.* (2009) *The Elements of Statistical learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer, New York.

Hindorff,L. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362.

Kolmogorov,V. and Zabih,R. (2004) What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 147–159.

Manolio,T. (2010) Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, **363**, 166–176.

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

Maronna,R. *et al.* (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

Mazumder,R. *et al.* (2011) SparseNet: Coordinate descent with non-convex penalties. *J. Am. Stat. Assoc.* (in press).

Meinshausen,N. and Buhlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B* , **72**, 417–473.

Rinaldo,A. (2009) Properties and refinements of the fused lasso. *Ann. Stat.*, **37**, 2922–2952.

Sabatti,C. *et al.* (2003) False discovery rate in linkage and association genome screens for complex disorders. *Genetics*, **164**, 829–833.

Schaid,D. (2004) Linkage disequilibrium testing when linkage phase is unknown. *Genetics*, **166**, 505–512.

Seaman,S. and Muller-Myhsok,B. (2005) Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.*, **76**, 399–408.

She,Y. (2009) Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Stat.*, **3**, 384–415.

Spencer,C. *et al.* (2011) Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS Genet.*, **7**, e1001337.

Storey,J. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

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

Tibshirani,R. and Wang,P. (2008) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, **9**, 18–29.

Tibshirani,R. *et al.* (2005) Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B* , **67**, 91–108.

Wan,X. *et al.* (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.

Wang,W. *et al.* (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.

Wehkamp,J. *et al.* (2004) NOD2 (CARD15) mutations in Crohn's disease are associated with diminished mucosal α-defensin expression. *Gut*, **53**, 1658–1664.

Wei,Z. *et al.* (2009) Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics*, **25**, 2802–2808.

Ye,G. and Xie,X. (2010) Split Bregman method for large scale fused Lasso. *Comput. Stat. Data Anal.*, **55**, 1552–1569.