Statistics and Its Interface Volume 3 (2010) 501–511

Simultaneous set-wise testing under dependence, with applications to genome-wide association studies

Wei Wang, Zhi Wei and Wenguang Sun*

We consider the problem of identifying disease-associated genomic regions in genome-wide association studies (GWAS). It is shown that conventional single SNP analysis can be greatly improved by (i) exploiting the spatial dependency and (ii) conducing set-wise analysis. The SNP set association problem can be conceptualized as the problem of simultaneously testing a large number of sets of hypotheses. We use hidden Markov models to exploit the linkage disequilibrium information in GWAS data, based on which a data-driven screening procedure (GLIS) is proposed. GLIS is shown to be optimal in the sense that it has the smallest missed set rate (MSR) among all valid false set rate (FSR) procedures. The numerical results demonstrate that the proposed procedure controls the FSR at the desired level, enjoys certain optimality properties and outperforms conventional combined p-value methods. We apply the GLIS procedure to analyze a Type 1 diabetes (T1D) GWAS dataset for detecting T1D associated genomic regions. The results show that our proposed SNP set analysis not only provides better biological insights, but also increases the statistical power by pooling information from different samples.

Keywords and phrases: Hidden Markov model, Large-scale multiple testing, Conjunction test, Partial conjunction test, Genome-wide association studies.

1. INTRODUCTION

Driven by advances in low-cost, high-throughput profiling technologies, genome-wide association studies (GWAS) have been widely used to interrogate the architecture of the whole human genome. GWAS have shown to be successful in detecting genetic variants that contribute to complex diseases. For example, GWAS have identified 53 new common genomic regions that are associated with autoimmune diseases [21], and have become predominant tools as a first step to localize the unknown weak variants.

In large-scale GWAS, it is typical to test hundreds of thousands of correlated markers simultaneously. However, conventional multiple testing procedures often suffer from low statistical power and lack of replicable findings, which have greatly limited the practical advantage of GWAS, especially in detecting markers with moderate or small effects. For example, recent comparative analyses of different GWAS on the same disease suggest that even for the most significant SNPs, the significance indicated by one study may not necessarily show up in another study. The low signal-to-noise ratio in most genotype data sets, a typical consequence of the "large p, small n" paradigm, provides new statistical challenges for developing a testing procedure with both high sensitivity and low error rate: a liberal p-value cutoff yields too many false positive results, yet a strict p-value cutoff tends to wipe out most interesting effects. In this article, we discuss two useful strategies that promise to improve the signal-to-noise ratio in GWAS data: utilization of spatial dependency and set-wise analysis. We shall develop an asymptotically optimal data-driven procedure in a compound decision theoretic framework where the grouping and dependency structures in the GWAS data can be exploited in a unified way.

During the meiosis process in a germ cell, a cross-over breaks parental chromosomes non-randomly into inheritable segments and then form gametes by recombining those segments. The single nucleotide polymorphisms (SNPs) within a segment will be inherited together with high probability and random combinations of all possible SNP states within the segment are prohibited. This co-segregation of adjacent SNPs results in the so-called linkage disequilibrium (LD). For a given SNP, its LD dependency with distant marker sites is attenuated over generations because of recombination. On the other hand, adjacent SNPs are likely to show strong association with each other. In general, the LD dependency decreases with the physical distance between two SNPs.

The linear block structure in the SNP data is very informative for constructing efficient testing procedures. First, if a SNP is disease-associated, then it is likely that the neighboring SNPs are also disease-associated (due to the co-segregation). Therefore, when deciding the significance level of a SNP, the neighboring SNPs should be taken into account. This shows that exploiting the dependency structure is very promising to increasing the efficiency of existing screening procedures. Second, the linear block structure

*Corresponding author.

indicates that set-wise analysis in general yields more interpretable scientific findings. The proposed SNP set analysis, resting upon the assumption that SNPs underlying a disease phenotype work in groups, may help to identify weak effects markers that are undetectable in single-SNP association studies. Setting a goal that target groups of important variables will significantly increase the screening power and reduce the number of false positives.

In summary, exploiting dependency and combining hypotheses in sets are two useful strategies to increase the signal to noise ratio in the sample, and are especially suitable for application in large-scale GWAS due to the special structures of the SNP data. A major goal of this research is to develop a multiple testing procedure that simultaneously incorporates the dependency and grouping information described above. Next we shall discuss some related works on multiple testing under dependence, then introduce existing methods for SNP set analysis, and finally outline the main ideas and advantages of our proposed research.

The correlation effects on multiple testing procedures have been extensively studied in the literature, see [14, 26, 31, 28, 10], among others. However, most previous research has been focused on the validity issue of different false discovery rate (FDR, Benjamini and Hochberg 1995 [4]) procedures. Although the dependency structure is highly informative, it has been largely ignored by most conventional methods [5, 14, 26, 38]. For example, the works by [5, 13, 38] showed that the FDR is controlled at the nominal level by the BH procedure under different dependence assumptions, supporting the "do nothing" approach which treats all dependent tests as if they were independent.

However, the dependency information is highly informative for constructing more efficient tests. Sun and Cai (2009) [33] derived an optimal procedure, based on the local index of significance (LIS), for multiple testing in a hidden Markov model (HMM). Numerical results demonstrated that the performance of conventional p-value thresholding procedures can be substantially improved by exploiting the HMM-dependency. The LIS procedure was recently applied to a GWAS of Type I diabetes (T1D) and compared with the BH procedure [37]. Significant improvement in rankings for T1D loci was achieved. For instance, a recent GWAS meta-analysis has confirmed 46 T1D susceptibility loci [2], among which 3 are on the top 500 list identified by BH and 7 are on the top 500 list identified by LIS (c.f. Table 2 in [37]). By exploiting dependency, the signal to noise ratio is greatly increased by integrating information from adjacent locations. The precision of tests is greatly improved in the sense that (1) the number of false positives is greatly reduced and (2) the statistical power to reject a non-null is substantially increased. This indicates that dependence can make the testing problem "easier" and is a *blessing* if efficiently utilized. A major goal of this article is to extend the LIS procedure in [33] for set-wise FDR analysis.

The strategy of conducting set-wise analyses in multiple comparisons is driven by both the need from scientific applications and the statistical consideration of screening power. First, conjunction analysis of SNP sets is of great biological interest, since groups of SNPs often serve as better proxies to capture disease association than any single SNP [9]. Second, recognizing that the "most significant SNPs" approach is less capable of separating the majority of truly associated signals from background noises, researchers have hypothesized that the "most significant SNP sets" approach, which jointly considers multiple SNPs in a genetic or biological meaningful set, might complement the "most significant SNPs" approach for analyzing data and interpreting results from such studies [35]. The proposed SNP set analysis, resting upon the assumption that SNPs underlying a disease phenotype work in groups, may help to identify weak-effects markers that are undetectable in single-SNP association studies. The set information can be derived from chromosome bands, biological pathways, functional terms (Gene Ontology), and etc (see Section 4 for more details). A SNP set analysis conducted in [35] successfully discovered a significant association between Crohn Disease and the IL12/IL23 pathway which harbors 20 genes. Such findings suggest that functional grouping information is very useful in that weak signals from individual observations can be pooled together to exhibit overall significance. However, the LD dependency information is ignored in [35] and the multiplicity issue is essentially not addressed.

In this article, we consider a new multiple testing approach for set-wise FDR analysis. Our approach addresses the multiplicity issue since it controls the false discovery rate asymptotically. In addition, by taking into account the LD dependency, our approach is expected to bring further improvements over the existing methods on SNP set analyses. As in [37], we use HMMs to model the SNP-SNP and SNP-trait associations. The fundamental difference is that the goal has changed to the identification of *groups of markers* or *SNP sets* that are associated with the disease. This goal can be achieved by testing the *conjunction* of null hypotheses and *partial conjunction* of null hypotheses [3]. We expect that the signal to noise ratio in the sample can be greatly enhanced by (i) integrating the information of all SNPs in a set and (ii) exploiting the spatial dependency.

The article is organized as follows. The hidden Markov model, theoretical framework for set-wise testing and a data-driven procedure are discussed in detail in Section 2. In Section 3, simulation studies are carried out to compare the numerical performances of our approach vs. conventional methods. In Section 4, our procedure is applied to the analysis of data from a GWAS of T1D for identifying disease-associated genomic regions. We conclude the article with a discussion of results and open problems.

2. SIMULTANEOUS ANALYSIS OF SETS OF CORRELATED HYPOTHESES

By combining all hypotheses in a set, we can form a new hypothesis at the set level; examples include the *conjunction* of null hypotheses (global null that all hypotheses are true), *disjunction* of null hypotheses (all hypotheses are false) and *partial conjunction* of null hypotheses [3]. Conventionally, Fisher's combined p-value method is the best known method for testing conjunction of null hypotheses and has been widely used in many applications, such as meta-analysis of microarray experiments and brain imaging studies [20, 39, 24, 16, 27, 19]. Simultaneous partial conjunction tests has recently been considered in FDR analysis by Benjamini and Heller (2008)[3].

In this section, we first introduce some important concepts and existing methods for simultaneous analysis of sets of hypotheses, then develop an "oracle" testing procedure in a compound decision theoretic framework by assuming that the correlation structure and distributional information are known. Finally, we discuss issues related to the practical implementation of the oracle procedure, including a hidden Markov model (HMM) for SNP data and a data-driven procedure that mimics the oracle procedure based on HMMs.

2.1 A decision-theoretic formulation for set-wise multiple testing

Suppose that we have divided the SNPs into K sets (see Section 4 for more details), and in set k there are m_k SNPs. Let $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{km_k})$ denote the unknown states in set k, where $\theta_{ki} = 1$ if SNP i from set k is disease associated and $\theta_{ki} = 0$ otherwise. Depending on an investigator's interest, a set-wise screening procedure claims a SNP set is interesting if (i) at least one SNP in the set is disease-associated; or (ii) at least u_k out of m_k SNPs in the set are disease-associated; or (iii) all m_k SNPs are disease associated. By convention, (i)–(iii) are respectively referred to as conjunction test, partial conjunction test and disjunction test. Conjunction and disjunction tests can be viewed as special cases of partial conjunction tests.

For a given SNP pattern of interest, we can define the null and non-null parameter spaces for $\boldsymbol{\theta}_k$. For example, the null space Θ_0^k for testing global null and partial conjunction are $\Theta_0^k = \{\boldsymbol{\theta}_k : \sum_{i=1}^{m_k} \theta_{ki} = 0\}$ and $\Theta_0^k = \{\boldsymbol{\theta}_k : \sum_{i=1}^{m_k} \theta_{ki} < u_k\}$, respectively. The non-null state spaces Θ_1^k can be obtained as the complement of the corresponding null spaces. For the K sets of hypotheses, define a binary vector

$$\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_K) \in \{0,1\}^K,$$

where

$$\vartheta_k = 0 \text{ if } \boldsymbol{\theta}_k \in \Theta_0^k \text{ and } \vartheta_k = 1 \text{ otherwise.} \tag{1}$$

Therefore the SNP set selection problem can be restated as the simultaneous testing of K new hypotheses: $\mathcal{H}_{k0} : \boldsymbol{\theta}_k \in \Theta_0^k$ versus $\mathcal{H}_{k1} : \boldsymbol{\theta}_k \in \Theta_1^k$, $k = 1, \ldots, K$, where set k is selected if the null hypothesis $\mathcal{H}_{k0}$ is rejected at the set level. Here, we are interested in inference of the unknown ϑ_k's based on the observed data and need to solve K component problems simultaneously. A solution to this problem can be represented by a compound decision rule

$$\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K) \in \{0,1\}^K,$$

where $\delta_k = 1$ if we claim that $\mathcal{H}_{k0}$ is false and $\delta_k = 0$ otherwise. Under this formulation, a false positive (negative) occurs if we decide $\delta_k = 1$ ($\delta_k = 0$) while $\vartheta_k = 0$ ($\vartheta_k = 1$). To summarize the set-wise testing results, we use the false set rate (FSR) and missed set rate (MSR) to combine the false positive and false negative results in set-wise testing:

$$\mathrm{FSR} = E\left\{\frac{\sum_{k=1}^K (1-\vartheta_k)\delta_k}{(\sum_{k=1}^K \delta_k) \vee 1}\right\} \tag{2}$$

and

$$\mathrm{MSR} = E\left\{\frac{\sum_{k=1}^K \vartheta_k(1-\delta_k)}{(\sum_{k=1}^K \vartheta_k) \vee 1}\right\}. \tag{3}$$

Specifically, the FSR is the expected proportion of falsely rejected sets among all rejections and the MSR is the expected proportion of non-null sets that are missed. The goal of set-wise multiple testing is to find a decision rule $\boldsymbol{\delta}$ that minimizes the MSR for a prespecified FSR level.

2.2 Combined p-value approach and its drawbacks

In this section, we introduce a screening procedure SBH based on combined p-values proposed by Benjamini and Heller (2008) [3]. In microarray data analysis, several methods have been introduced for testing the significance of multiple gene sets [25, 12]. However, few of these methods utilize the dependency among individual tests within a set. We shall see that the SBH procedure is inefficient and can be greatly improved by our GLIS procedure which exploits the available spatial dependency.

Suppose there are n_1 cases and n_2 controls being genotyped over the m_k SNPs in set k, $k = 1, \ldots, K$. The total number of SNPs is $m = \sum_{k=1}^K m_k$. We conduct a χ^2-test with 1 degree of freedom for each SNP to assess the association between the allele frequencies and the disease status; the p-values from the χ^2-tests are then recorded. Let $p_{(1)}^k, \ldots, p_{(m_k)}^k$ be the ordered p-values from the kth set. Denote by H_{u/m_k}^k the partial conjunction of hypotheses that at least u out of m_k hypotheses in set k are false. The Simes' p-value can be used to summarize the p-values from set k into a single index

$$p_{u/m_k}^k = \min_{j=1,\ldots,m_k-u+1}\left\{\frac{m_k-u+1}{j} p_{(u-1+j)}^k\right\}. \tag{4}$$

Then H_{u/m_k}^k is rejected if p_{u/m_k}^k is small.

Benjamini and Heller [3] proposed a two-stage procedure for FSR control. In the first stage, Simes' p-values are obtained for all sets. In the second stage, the p-values are ordered as $p^{(k)}_{u/m_k}, k = 1, \ldots, K$. Denote by $H^{(k)}_{u/m_k}$, $k = 1, \ldots, K$, the corresponding hypotheses. The BH step-up procedure [4] is then applied to the ordered p-values to determine which sets should be rejected:

$$\text{Let } l = \max\{k : p^{(k)}_{u/m_k} \le \frac{k}{K}\alpha\},$$
$$\text{then reject } H^{(k)}_{u/m_k},\ k = 1, \ldots, l.$$

It was shown in [3] that this procedure, referred to as the Simes-BH (SBH) procedure, controls the FSR at the nominal level α. In addition, it was shown that the SBH procedure is still valid under different dependency assumptions.

However, the SBH procedure is highly inefficient because the information of the dependence structure among SNPs in a set can be exploited to construct more efficient tests. Next we develop an oracle procedure for testing sets of correlated hypotheses. In the derivation, it is assumed that the correlation structure and distributional information are known. A data-driven procedure based on an HMM that mimics the oracle procedure will be discussed in Section 2.5 for situations where such information is unknown.

2.3 The oracle procedure for FSR control

To facilitate the future implementation of our oracle procedure, we shall use z-values instead of p-values for methodological development. Specifically, given the two-sided p-value and odds ratio of a SNP, we can convert a p-value to a z-value using the following transformation

$$z_i = \begin{cases} \Phi^{-1}(1 - \frac{p_i}{2}) & \text{if odds ratio} > 1 \\ \Phi^{-1}(\frac{p_i}{2}) & \text{otherwise} \end{cases}. \tag{5}$$

It is reasonable to assume that a z-value is distributed as $N(0, 1)$ under the null, and is distributed as a normal mixture under the alternative. The normal mixture model is a dense class, which is general enough to approximate almost all mixture distributions and has been found in a wide range of applications. Such a transformation also greatly facilitates the implementation of our oracle procedure because various methods for consistently estimating the normal mixtures have been developed in the literature. Details on model parameter estimation (EM algorithm) in a normal mixture will be discussed in Section 2.5.

Now we turn to the derivation of our oracle procedure for set-wise testing of correlated hypotheses. The basic assumptions on the data structure are (i) the data for each hypothesis has been summarized to a z-value (based on a χ^2-test), (ii) the z-values are correlated and the correlation structure is known. We assume the conditional distributions for z-values

$$z_{ki}|\theta_{ki} \sim (1 - \theta_{ki})F_0 + \theta_{ki}F_1. \tag{6}$$

where F_0 and F_1 are the null and non-null distribution functions, respectively. The corresponding density functions are denoted by f_0 and f_1. When z-value is used, F_0 is standard normal $N(0, 1)$ and F_1 is a normal mixture $F_1 = \sum_{l=1}^{L} N(\mu_l, \sigma_l^2)$, where L is the number of components in the mixture. In this section, we assume that there is an oracle that knows F_1 and the model that describes the dependency of the tested hypotheses with parameter Ψ. The specific definition of Ψ and the situation where such information is unknown will be considered later. Next we study the oracle's response to the set-wise simultaneous testing problem.

Consider a general decision rule $\boldsymbol{\delta} = (\delta_k : k = 1, \ldots, K) \in \{0, 1\}^K$, where $\delta_k = 1$ indicates that SNP set k is rejected and $\delta_k = 0$ otherwise. The multiple testing problem is closely related to a weighted classification problems. Specifically, let λ be the known relative cost of a false positive to a false negative. Define a weighted classification problem with loss function

$$L_\lambda(\boldsymbol{\vartheta}, \boldsymbol{\delta}) = \frac{1}{K}\sum_{k=1}^{K}\{\lambda(1 - \vartheta_k)\delta_k + \vartheta_k(1 - \delta_k)\}. \tag{7}$$

It was shown in Sun and Cai (2007, 2009) [32, 33] that the multiple testing and weighted classification problems are essentially equivalent under a monotone ratio condition (MRC). Specifically, let $\mathcal{T}$ be the collection of all test statistics that satisfy the MRC and $\mathcal{D}_\alpha$ the collection of the testing rules at FSR level α and of the form $\boldsymbol{\delta} = I(\boldsymbol{T} < c\mathbf{1})$. Suppose that the classification risk with the loss function defined in (7) is minimized by $\delta^\lambda\{\boldsymbol{T}, c(\lambda)\}$, so that $\boldsymbol{T}$ is the optimal statistic in the weighted classification problem. If $\boldsymbol{T} \in \mathcal{T}$, then $\boldsymbol{T}$ is also optimal in the multiple testing problem, in the sense that for each FSR level α, there exists a unique $c(\alpha)$ such that $\delta^\alpha\{\boldsymbol{T}, c(\alpha)\}$ controls the FSR at level α with the smallest MSR level in $\mathcal{D}_\alpha$.

It can be shown that the *optimal* classification rule that minimizes the classification risk is the Bayes rule

$$\boldsymbol{\delta}\{\boldsymbol{\Lambda}, (1/\lambda)\mathbf{1}\} = (\delta_k : k = 1, \ldots, K),$$

where

$$\Lambda_k = P_\Psi(\vartheta_k = 0|\boldsymbol{z})/P_\Psi(\vartheta_k = 1|\boldsymbol{z})$$

and $\delta_k = I(\Lambda_k < 1/\lambda)$. The equivalence between multiple testing and weighted classification under dependence [33] implies that $\boldsymbol{\Lambda}$ is also the optimal test statistic for FSR control. Next we define the *generalized local index of significance* (GLIS)

$$\text{GLIS}_k = P_\Psi(\vartheta_k = 0|\boldsymbol{z}),$$

for $k = 1, \ldots, K$. Note that $\text{GLIS}_k = \Lambda_k/(1 + \Lambda_k)$ is strictly increasing in Λ_k, the (oracle) optimal multiple testing procedure must be of the form

$$\boldsymbol{\delta}(\mathbf{GLIS}, c_{OR}\mathbf{1}) = [I(\text{GLIS}_k < c_{OR}) : k = 1, \ldots, K], \tag{8}$$

where the oracle cutoff c_{OR} is given by

$$c_{OR} = \sup\{c \in (0,1) : \mathrm{FSR}(\mathbf{GLIS}, c\mathbf{1}) \le \alpha\}.$$

The oracle procedure (8) provides a benchmark for developing and evaluating different FSR procedures. The difficulty of the implementation of oracle procedure (8) varies according to the model specification. In Section 2.4, we introduce an HMM for SNP data, then discuss the implementation of GLIS oracle procedure in Section 2.5.

2.4 A hidden Markov model for SNP data

The hidden Markov Model (HMM) is a classical model to capture linear dependency and, due to DNA's linear primary structure, it has been widely applied to analyze genomic data, such as inferring protein binding sites from ChIP-Chip data [23] and detecting DNA number copy alternations from array CGH data [15]. An optimal testing procedure under HMM-dependency was developed in Sun and Cai (2009) [33]. For single SNP analysis, we have implemented HMMs to characterize the dependency among neighboring SNPs [37] and successfully generalized Sun and Cai's procedure to model multiple heterogeneous HMMs.

In an HMM, it is assumed that each SNP in the chromosome has two possible hidden states: disease-associated or non-disease-associated, and the states of all SNPs along the chromosome form a Markov chain. In our application, the observed genotype data are assumed to be generated conditional on the hidden states via a normal mixture model. Specifically, let $\boldsymbol{\theta}_k = \{\theta_{k1}, \ldots, \theta_{km_k}\}$ denote the underlying states of the SNP sequence in set k, where $\theta_{ki} = 1$ indicates that SNP i from set k is disease-associated and $\theta_{ki} = 0$ otherwise. In the current section, we only consider testing on one chromosome; the multiple-chromosome situation will be considered in Section 2.6. Assume that $\boldsymbol{\theta} = (\boldsymbol{\theta}_k : k = 1, \ldots, K)$ is distributed as a stationary Markov chain with transition probability

$$a_{ij} = P(\theta_s = j | \theta_{s-1} = i) \tag{9}$$

In an HMM, the observed z-values are assumed to be conditionally independent given the hidden states:

$$P(\boldsymbol{z}_k | \boldsymbol{\theta}_k, \mathcal{F}) = \prod_{i=1}^{m_k} P(z_{ki} | \theta_{ki}, \mathcal{F}), \tag{10}$$

for $k = 1, \ldots, K$. Denote by $\mathcal{A} = (a_{ij})$ the transition matrix, $\boldsymbol{\pi} = (\pi_0, \pi_1)$ the stationary distribution, $\mathcal{F} = \{F_0, F_1\}$ the observation distribution, and $\Psi = (\mathcal{A}, \boldsymbol{\pi}, \mathcal{F})$ the collection of all HMM parameters.

2.5 A data-driven procedure

The optimal testing procedure (8) is difficult to implement because it is hard to determine the optimal cutoff c_{OR} directly. Also, in practice, the HMM parameters Ψ are unknown. As before, we first obtain the estimated parameters $\hat{\Psi}$ by the EM algorithm, then plug-in $\hat{\Psi}$ to obtain

$$\widehat{\mathrm{GLIS}}_k = P_{\hat{\Psi}}(\vartheta_k = 0 | \boldsymbol{z}).$$

The details of the EM algorithm for a normal mixture model are given in [33]. In situations where the number of components L is unknown, we can use the BIC criterion to determine the best choice of L. Specifically,

$$\mathrm{BIC} = \log\{P(\hat{\Psi}_L | \boldsymbol{z})\} - \frac{|\Psi_L|}{2} \log(m),$$

where $P(\Psi_L | \boldsymbol{z})$ is the likelihood function, $\hat{\Psi}_L$ is the maximum likelihood estimate of HMM parameters, and $|\Psi_L|$ is the total number of HMM parameters.

The next step is to rank the plug-in GLIS statistics from all sets and choose an appropriate cutoff. Denote by $\widehat{\mathrm{GLIS}}_{(1)}, \ldots, \widehat{\mathrm{GLIS}}_{(K)}$ the ranked plug-in values and $H_{(1)}, \ldots, H_{(K)}$ the corresponding hypotheses. In light of the oracle procedure, we propose the following data-driven procedure (GLIS):

$$\text{Let } l = \max\left\{ k : \frac{1}{k} \sum_{j=1}^{k} \widehat{\mathrm{GLIS}}_{(j)} \le \alpha \right\}, \tag{11}$$

then reject all $H_{(k)}$, $k = 1, \ldots, l$. Given the estimated parameters $\hat{\Psi}$, the GLIS statistic

$$P\left(\sum_{i=1}^{m_k} \theta_{ki} < \varepsilon | \boldsymbol{z} \right)$$

can be computed for any partial conjunction patterns, namely, no more than ε non-nulls in the set. Specifically, the calculation involves exhaustively enumerating all possible patterns and summing up their probabilities based on the forward-backward algorithm [29]. The choice of ε is based on prior genetics knowledge and investigator's experiences, e.g., $\varepsilon = 4$, if we believe that for a true disease variant, its two neighbors on each side (dependent on markers' density) shall be LD dependent and show significant association; otherwise it may be a false positive due to noise.

Compared to the combined p-value procedure SBH, the advantages of the GLIS include: (i) Interpretability. By definition, the GLIS can be interpreted as the probability of a genomic region having the spatial pattern of interest given the observed genotypic data. In contrast, the meaning of the combined p-value, obtained via a step-up procedure, is not obvious. (ii) Accuracy. The numerical results in our simulation studies show that the GLIS approach can accurately achieve the nominal FSR level, whereas the SBH procedure is over-conservative. (iv). Efficiency. By exploiting the spatial correlation, the GLIS produces more efficient rankings

of genomic regions than the Simes' p-value. Hence it is expected that the GLIS identifies more true signals than the SBH procedure at the same FSR cost.

The following theorem, which can be proved similarly to the Theorems 5 and 6 in [33], shows that the data-driven procedure is asymptotically valid and optimal.

Theorem. *Consider the HMM defined by (9) and (10). Let $\widehat{\Psi}$ be a consistent estimate of the HMM parameters. Denote by $\widehat{GLIS}_{(1)},\ldots,\widehat{GLIS}_{(K)}$ the ranked plug-in values, and $H_{(1)},\ldots,H_{(m)}$ the corresponding hypotheses. Then under the regularity conditions (1)–(5) in [33], the data-driven procedure (11) controls the FSR at level α asymptotically. In addition, let MSR_{OR} and MSR_{DD} be the MSR levels of the oracle procedure (8) and the data-driven procedure (11), respectively, then $MSR_{DD} = MSR_{OR} + o(1)$.*

2.6 Pooled analysis with multiple chromosomes

The chromosomes in a genome segregate independently and may exhibit different dependency structures. It is therefore more appropriate to model each chromosome separately. In this section, we generalize the previous data-driven procedure for multiple-chromosome analysis. The key issue is how to combine the simultaneous inferences made for separate chromosomes to achieve optimal genome-wise FSR control. See [6] for more theoretical backgrounds on the optimality of multiple testing with groups.

A straightforward approach to combining the analyses from different chromosomes is the so-called *separate* analysis [11], which suggests applying an inference procedure to each chromosome at the same test level. However, this approach is not optimal in general. Let $\boldsymbol{z}^c$ be the observed data for chromosome c, $c = 1,\ldots,C$. Denote by $\widehat{\Psi^c}$ the estimated HMM parameters for chromosome c and $\widehat{\text{GLIS}}_k^c$ the corresponding test statistics. In light of the data-driven CLfdr procedure in [6], we propose the following *pooled* procedure for multiple-chromosome set-wise analysis:

Step 1. Calculate the plug-in GLIS statistic $\widehat{\text{GLIS}}_k^c = P_{\Psi^c}(\vartheta_k^c = 0|\boldsymbol{z}^c)$ for individual chromosomes $c = 1,\ldots,C$.

Step 2. Combine and rank the plug-in GLIS statistic from all chromosomes. Denote by $\widehat{\text{GLIS}}_{(1)},\ldots,\widehat{\text{GLIS}}_{(\sum_{c=1}^C K^c)}$ the ordered values and $H_{(1)},\ldots,H_{(\sum_{c=1}^C K^c)}$ the corresponding hypotheses.

Step 3. Reject all $H_{(i)}$, $i = 1,\ldots,l$, where $l = \max\{i : (1/i)\sum_{j=1}^i \widehat{\text{GLIS}}_{(j)} \le \alpha\}$.

It can be shown that this pooled procedure is optimal in the sense that the genome-wise MSR is minimized subject to a constraint on the genome-wise FSR. One important feature of this pooled procedure is that different chromosome-wise FSR levels are chosen such that genome-wise MSR level is minimized. The 3-step procedure is a hybrid strategy that has combined features from both pooled and separate analyses. Specifically, it is a "separate" analysis because, in step 1, the grouping information is exploited to calculate chromosome-wise HMM parameters; it is also a "pooled" strategy because, in steps 2 and 3, the group labels are dropped and the rankings of all hypotheses are determined globally. The difference between our approach and Efron's approach is that we suggest a different way on how the simultaneous inferences from different chromosomes may be combined: Efron suggests using identical test levels for all chromosomes, whereas we suggest using different test levels for different chromosomes. Again, we refer to [6] for more insights and discussions on the advantages of our pooled strategy.

3. SIMULATION STUDIES

In this section, we conduct simulation studies to compare the numerical performances of the GLIS procedure vs. the SBH procedure. The SBH procedure is inefficient because the dependency structure is ignored. We will show that GLIS improves SBH by exploiting the dependency information among adjacent SNPs.

As an illustrative example, we consider 2 chromosomes, each with 2,000 SNPs. We generate two Markov chains $\boldsymbol{\theta}^c = (\theta_i^c)_{i=1}^{2000}$, $c = 1, 2$, with transition matrices $\mathcal{A}^1 = (0.98, 0.02; 0.3, 0.7)$ and $\mathcal{A}^2 = (0.98, 0.02; 0.05, 0.95)$, respectively. Conditional on the hidden states θ_i^c, the observations z_i^c are generated as

$$z_i^c|\theta_i^c \sim (1-\theta_i^c)N(0,1) + \theta_i^c N(\mu_c, 1).$$

Next, we define the set size to be 20, namely, from the first SNP on a chromosome, every 20 consecutive SNPs are grouped as a set. Consequently, we have 200 sets in total for the two chromosomes. Let the partial conjunction pattern threshold $\varepsilon = 5$, i.e., we have our partial conjunction null $\Theta_0 = \{\boldsymbol{\theta}_k : \sum_{i=1}^{20}\theta_{ki} < 5\}$. Our goal is to find the SNP sets with the pattern of our interest while controlling the FSR at a pre-specified level for the whole genome (combining chromosomes 1 and 2). For 200 replications, we apply GLIS and SBH procedures at FSR level $\alpha = 0.1$. We vary μ_1 from 1 to 4 with an increment 0.5 and $\mu_2 = \mu_1 + 1$, and plot the FSR and MSR levels as functions of μ_1.

The simulation results are shown in Figure 1. Panel (a) indicates that both procedures control the FSR at the nominal level $\alpha = 0.1$. However, GLIS gives a precise control of the FSR while SBH is over conservative. From Panel (b), we can see that the MSR of GLIS is much lower than that of SBH. Note that the smaller the value of μ, the weaker the signal, we conclude that the improvement in MSR levels brought by GLIS becomes larger as the signal is weaker. Additional simulation results indicate that the efficiency gain of GLIS is larger when the dependency is stronger. We omit

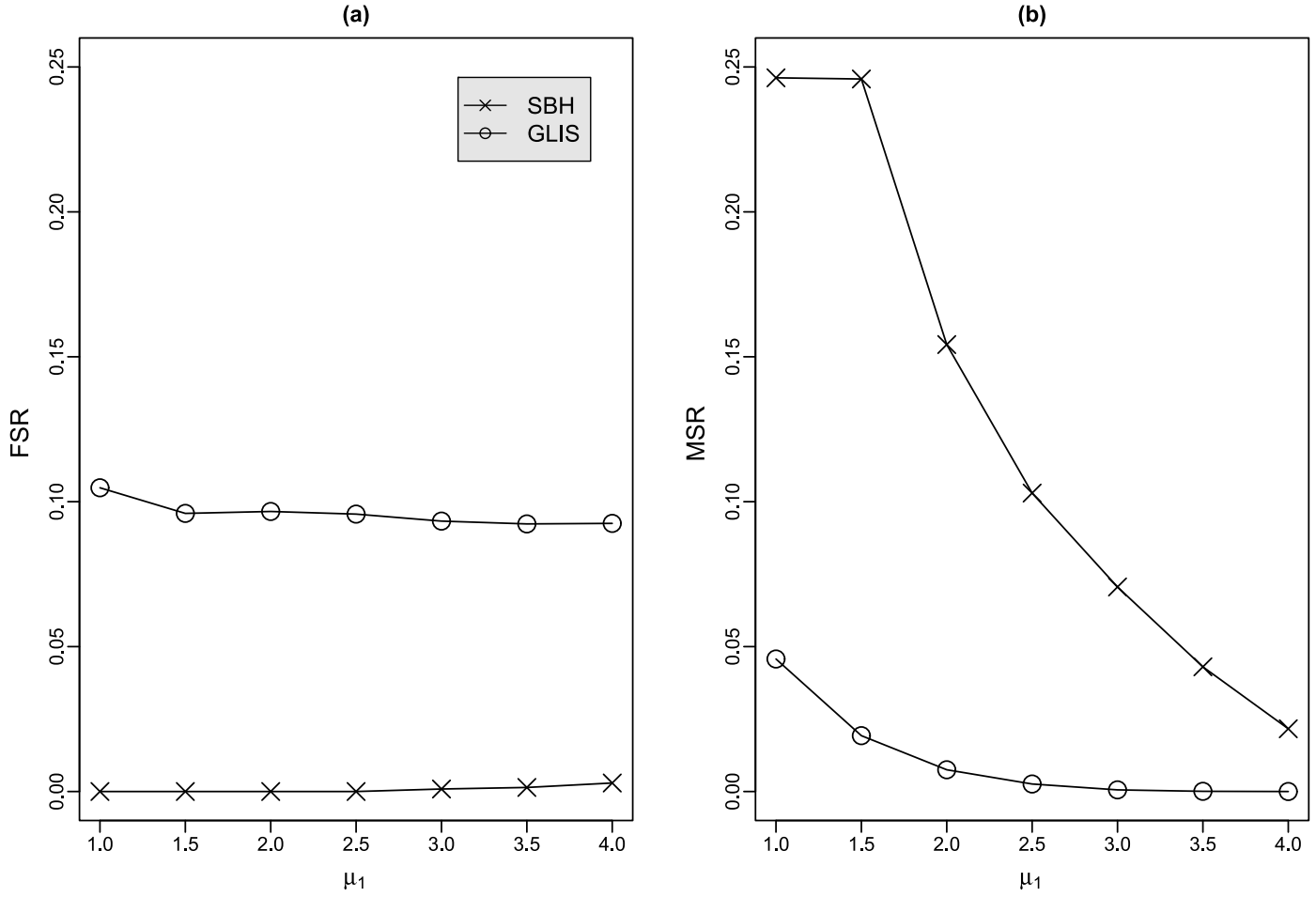

Figure 1. (a) FSR levels (b) MSR levels.

the details here and refer to [33] for more discussions on this issue.

It may be argued that the higher power of GLIS is gained at the price of a higher FSR level. Actually, another important source for the efficiency gain is in the improved rankings produced by GLIS. To illustrate, we plot Figure 2 to compare the ranking efficiencies of GLIS versus SBH. Here, the sensitivity is calculated as the average proportions of correctly identified SNP sets over the 200 replications. Using different significance thresholds, we calculate corresponding FSRs and sensitivities. We can see that under the same FSR level, GLIS discovers more true disease-associated SNP sets than SBH. Again, we observe the improvement is more dramatic when signals are weak. This makes GLIS particularly attractive to GWAS in finding genetics variants with moderate or small effects, which would be missed otherwise by conventional p-value based approaches.

From the above simulation studies we can see that, by modeling genomic dependency, the GLIS procedure can greatly improve the efficiency of detecting (weak) signals. The signal to noise ratio in the sample is increased by integrating information from adjacent SNPs. As a result, we may simultaneously reduce the number of false positives and increase the statistical power to reject a non-null. This confirms that dependence can make the testing problem "easier" and is a *blessing* if incorporated properly in a testing procedure [30, 33, 37].

Another issue is the efficiency of set-wise analysis versus single SNP analysis. Conceptually, the SNP set analysis is also more powerful because the signal to noise ratio is further enhanced by pooling information from all SNPs in a set. However, it is difficult to conduct a fair comparison of them because the goals of the two types of analyses are fundamentally different. Instead, we shall illustrate the differences between these two approaches in our real data analysis presented next.

4. APPLICATION TO T1D DATASETS

Childhood diabetes, also called Type 1 diabetes (T1D), is a multifactorial, autoimmune complex disease. Different from Type 2 diabetes, childhood diabetes is typically found in young individuals with onset as early as one year old and most cases are diagnosed before the age of 18. As a result,

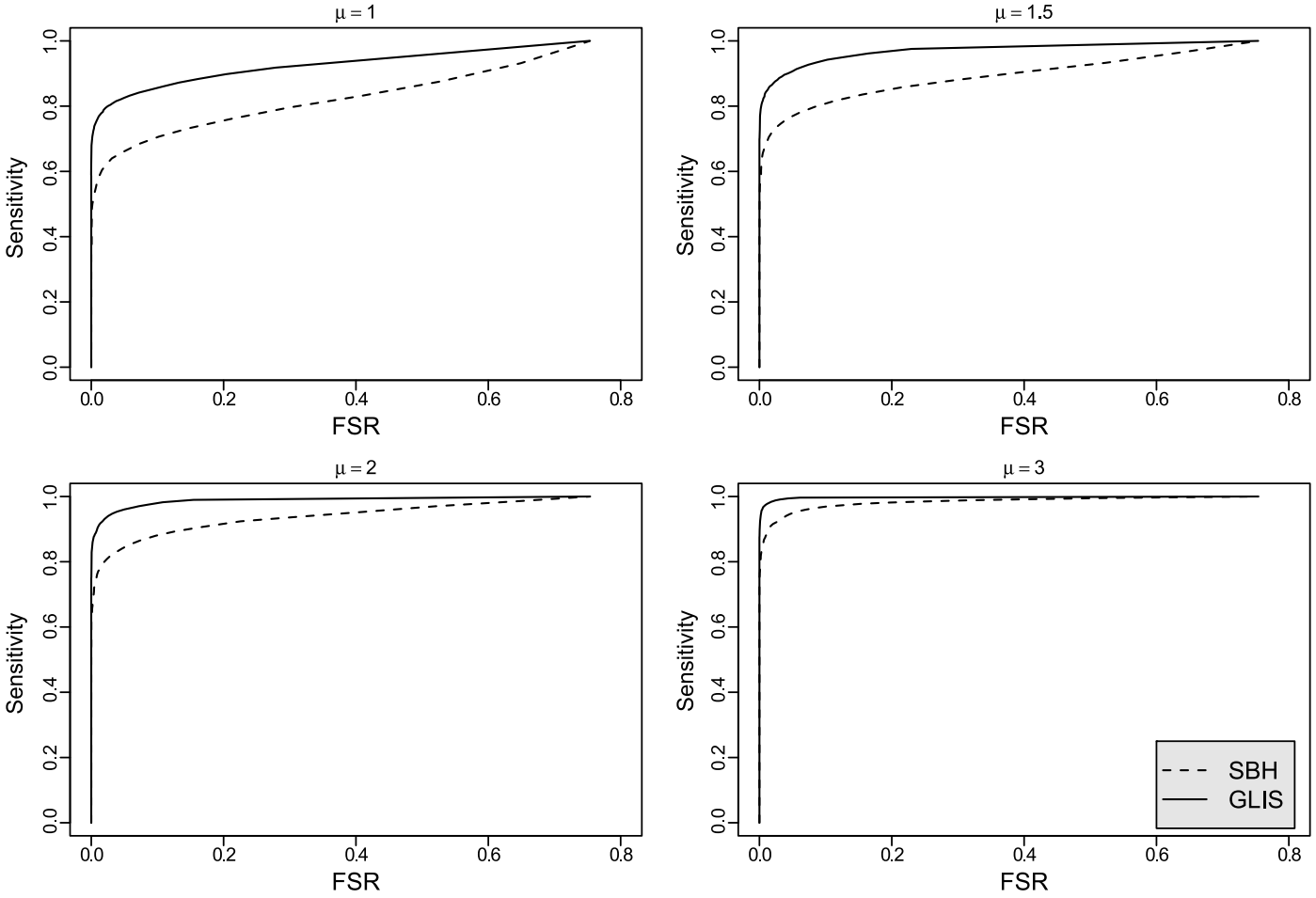

Figure 2. At the same FSR level, GLIS has higher sensitivity than SBH, especially when signals are weaker.

environmental factors complicate the childhood diabetes genetic analysis less than other complex diseases. Associations of six loci with child diabetes have been convincingly established because of their large effects. However, the established genetic associations with childhood diabetes only explain little more than 50% of the genetic risk. Many more genes with moderate or small effects remain to be discovered [34, 18]. Finding these unknown "weak" genes is the main target of GWAS.

The Wellcome Trust Case-Control Consortium (WTCCC), established in 2005, consists of a large number of research groups across the UK. One of the WTCCC aims was to explore the utility, design and analysis of GWAS for detecting genetic variants associated with most common diseases. In 2007, the WTCCC published data on 14,000 cases of seven common diseases and 3,000 shared controls [8]. All samples are genotyped by Affymetrix Mapping 500K arrays. As a real case study, we apply our proposed GLIS procedure to analyze the 2,000 T1D cases and the 3,000 shared controls. We perform a series of standard quality control procedures to eliminate markers with minor allele frequency less than 1%, Hardy-Weinberg Equilibrium p-values lower than 1e-6, or genotype no-call rate higher than 5%. In addition, we remove the problematic samples as specified in the WTCCC website. After the quality control, 397,780 SNPs from 1,963 T1D cases and 2,938 shared controls are eligible for further analysis.

4.1 Model selection and the estimation of HMM parameters

We first conduct a χ^2-test with d.f. of 1 for each SNP to assess the association between the allele frequencies and the disease status. Then we obtain z-values from the p-values using the transformations defined by (5). Each chromosome is modeled separately to obtain chromosome-specific HMM parameters Ψ^c. We assume that the null distribution is standard normal $N(0,1)$ and the non-null distribution is a normal mixture $\sum_{l=1}^{L} c_l N(\mu_l, \sigma_l^2)$. The number of components L in the non-null distribution is determined by the BIC criterion,

$$\text{BIC} = \log\{P(\hat{\Psi}_L^c | \boldsymbol{z}^c)\} - \frac{|\Psi_L^c|}{2}\log(m^c),$$

where $P(\Psi_L^c|\boldsymbol{z})$ is the likelihood function, $\hat{\Psi}_L^c$, the MLE of HMM parameters, $|\Psi_L^c|$, the number of HMM parameters and m^c, the number of SNPs for chromosome c.

4.2 Results for set-wise analysis

We group every 20 consecutive SNPs as a set. Our goal is to identify T1D associated genomic regions spanned by the 20 SNPs, which covers at least one T1D variant. Given the estimated parameters $\hat{\Psi}$, we calculate the GLIS statistics for the partial conjunction pattern based on

$$P\left(\sum_{i=1}^{m_k}\theta_{ki} < \varepsilon|\boldsymbol{z}\right).$$

For a true disease variant, whether genotyped or not by the Affymetrix Mapping 500K array, we expect its two genotyped neighbors on each side be LD dependent and show significant association. Therefore, we set $\varepsilon = 4$. We also try $\varepsilon = 2, 3, 5$ and see how sensitive the set analysis is to the choice of ε. We apply both the SBH and GLIS procedures at FSR level 0.001.

Note that here we simply group every 20 consecutive SNPs as a set. In practice, the grouping strategy can be improved in several ways by integrating various domain knowledge. First, we can divide SNPs into blocks based on their LD dependency information derived from Hapmap [7]. Second, we may group SNPs in the same haplotype block [17] as a set. Many tools are available for haplotype reconstruction, such as HAPLORE [41] and HapBlock [40]. Third, we can map SNPs to the genes they belong to. Then as in many pathway-based analyses [36, 35], genes, thus the affiliated SNPs, can be grouped as a set if they come from the same biological pathway or have the same molecular function. Finally, clustering methods may also be employed for grouping SNPs.

A meta-analysis based on recent GWAS has confirmed 46 genetic variants associated with T1D [2]. The numbers of non-NULLs claimed by the two procedures under different ε values are shown in Table 1. The numbers of the known T1D variants among the claimed non-NULLs are given in brackets in the same table. We can see that SBH is very conservative and claims much fewer non-NULLs than GLIS under the same FSR level. Accordingly, GLIS identifies more known T1D variants. As the threshold ε increases, both procedures tend to claim fewer non-NULLs. This is not surprising because fewer significant sets are expected when we require more significant individual signals to claim one set to be significant. Compared with the least stringent threshold $\varepsilon = 2$, the threshold $\varepsilon = 4$ or 5 seems to be a good tradeoff for this T1D GWAS dataset, which claims about hundreds fewer non-NULLs with only 4 fewer known T1D variants. For different applications, practitioners may choose an appropriate ε value based on their domain experience and knowledge. Table 2 lists the top 6 known T1D loci identified by GLIS using the threshold $\varepsilon = 4$. We can see that the ranking by GLIS statistics is quite different from that by SBH procedure. A good example is the ranking of SNP rs6441961 and SNP rs7221109. SBH places SNP rs6441961 *after* SNP rs7221109, with rank 157 and 125, respectively, while GLIS places SNP rs6441961 *before* SNP rs7221109, with rank 102 and 109, respectively, both improved though.

Table 1. The number of non-NULLs (known T1D variants)

	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$
GLIS	1714(19)	1562(18)	1394(15)	1203(15)
SBH	66(2)	60(2)	52(1)	47(1)

5. CONCLUSION AND DISCUSSION

This article develops HMM-based partial conjunction testing procedures for identifying disease associated genomic regions in analysis of large-scale GWAS data. The proposed GLIS procedure is extended from [33, 37] for SNP sets analysis. We show that dependency information, if taken into account, shall again bring improvement in the multiple testing of partial conjunction hypotheses as in multiple testing of single hypotheses. The numerical performances of our GLIS procedure are investigated using both simulated and real data. Compared to the SBH procedure [3] which ignores dependency information, our GLIS procedure is more powerful in identifying small to moderate signals.

We group adjacent SNPs in LD into a set. The motivation is that multilocus methods such as Haplotype-based analysis have been generally appreciated for their higher potential of detecting disease susceptibility region than do single-marker methods [1]. In addition, previous research found that some groups of SNPs serve as better proxies for the

Table 2. The top 6 known T1D susceptibility loci identified by GLIS

Chromosome	SNP	Position (Base)	GLIS	SBH	GLIS Rank	SBH Rank
6	rs9268645	32516505	0	3.43E-43	1	7
1	rs6679677	114015850	1.85E-49	4.34E-09	37	33
16	rs12708716	11087374	8.51E-25	5.70E-05	68	64
12	rs1265566	110179096	1.74E-24	0.000687	70	72
3	rs6441961	46327388	3.37E-13	0.014597	102	157
17	rs7221109	36023812	1.30E-12	0.007472	109	125

hidden SNPs than does any single SNP [9]. The improved efficiency brought by our method can be attributed to the utilization of the group structure and the modeling of the LD dependency of SNPs by HMM. For a set of SNPs, there may be two scenarios: (1) many of them show weaker individual disease association because of high LD; and (2) only one or few of them show strong disease association and the others show no association due to low LD. It should be noted that if a relatively high threshold for ε is set, for example, the suggested value 4 or 5, our proposed method may prefer (1) and miss those strong isolated SNPs surrounded by low LD dependent neighbors. Because of the complex genome heterogeneity, it may be hard to adjust the threshold values to be region specific. As a result, our proposed method, like many other multilocus methods, does not intend to replace the current single SNP mapping methods but serves as a useful complement.

The GLIS procedure can be improved in several ways. First, a discrete time HMM is used in the proposed method. However the SNPs genotyped on an array are not distributed with equal distance. An immediate improvement is to incorporate the between-SNP distances and use an inhomogeneous HMM to model the dependency. Second, correlations among SNPs are much more complicated and the LD dependency may not always decrease monotonically with the physical distance between two SNPs. It could be that pairs of SNPs that are tens of kilobases apart are in "complete" LD, whereas nearby pairs of SNPs from the same region are in weak LD. It is shown that a network is a more precise description of the complex SNP dependency [22]. It would be of interest to generalize our partial conjunction testing procedure from a Markov chain to a Markov random field. Third, it might be a strong assumption that the whole chromosome follow a stationary Markov chain. Instead of assuming homogeneous transition probabilities for the whole Markov chain, we may use different transition probabilities for different genomic regions and introduce a hierarchical Bayes model to model them. Such hierarchical Bayes models relax the homogeneous dependency assumption and may lead to better inference procedures. Finally, each candidate set has the same number of SNPs. When applying GLIS to GWAS, the practitioner can allow variable set sizes and determine an appropriate number of SNPs to include into a set based on their domain knowledge, e.g., LD block structure derived from HapMap. We expect that GLIS can be further improved by integrating such domain knowledge.

ACKNOWLEDGEMENTS

We thank the two referees for their helpful suggestions. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

Received 28 January 2010

REFERENCES

[1] Akey, J., Jin, L., and Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* **9**, 4 (Apr), 291–300. http://dx.doi.org/10.1038/sj.ejhg.5200619.

[2] Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D. J., Stevens, H., Todd, J. A., Walker, N. M., Rich, S. S., and Consortium, T. T. . D. G. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet 41*, 703–707.

[3] Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64**, 4 (Dec), 1215–1222. http://dx.doi.org/10.1111/j.1541-0420.2007.00984.x.

[4] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 1, 289–300. http://dx.doi.org/10.2307/2346101. MR1325392

[5] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*, 1165–1188. MR1869245

[6] Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc. 104*, 1467–1481. MR2597000

[7] Consortium, I. H. (2007a). A second generation human haplotype map of over 3.1 million snps. *Nature* **449**, 7164 (Oct), 851–861. http://dx.doi.org/10.1038/nature06258.

[8] Consortium, W. T. C. C. (2007b). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 7145 (Jun), 661–678.

[9] de Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat Genet* **37**, 11 (Nov), 1217–1223. http://dx.doi.org/10.1038/ng1669.

[10] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc.* **102**, 477, 93–103. MR2293302

[11] Efron, B. (2008). Simultaneous inference: when should hypothesis testing problems be combined? *Ann. Appl. Stat.* **2**, 1, 197–223. http://dx.doi.org/10.1214/07-AOAS141. MR2415600

[12] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Stat.* **1**, 1, 107–129. http://dx.doi.org/10.1214/07-AOAS101. MR2393843

[13] Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scand. J. Stat.* **34**, 2, 275–297. MR2346640

[14] Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Stat.* **30**, 1, 220–238. MR1892662

[15] Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal.* **90**, 1, 132–153. http://dx.doi.org/10.1016/j.jmva.2004.02.008. MR2064939

[16] Friston, K. J., Penny, W. D., and Glaser, D. E. (2005). Conjunction revisited. *Neuroimage* **25**, 3 (Apr), 661–667. http://dx.doi.org/10.1016/j.neuroimage.2005.01.013.

[17] Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**, 5576 (Jun), 2225–2229. http://dx.doi.org/10.1126/science.1069424.

[18] Hakonarson, H., Grant, S. F. A., Bradfield, J. P., Marchand, L., Kim, C. E., Glessner, J. T., Grabs, R., Casalunovo, T., Taback, S. P., Frackelton, E. C., Lawson, M. L., Robinson, L. J., Skraban, R., Lu, Y., Chiavacci, R. M., Stanley, C. A., Kirsch, S. E., Rappaport, E. F., Orange, J. S., Monos, D. S., Devoto, M., Qu, H.-Q., and Polychronakos, C. (2007). A genome-wide association study identifies kiaa0350 as a type 1 diabetes gene. *Nature* **448**, 7153 (Aug), 591–594. http://dx.doi.org/10.1038/nature06010.

[19] Heller, R., Stanley, D., Yekutieli, D., Rubin, N., and Benjamini, Y. (2006). Cluster-based analysis of fmri data. *Neuroimage* **33**, 2 (Nov), 599–608. http://dx.doi.org/10.1016/j.neuroimage.2006.04.233.

[20] Lazar, N. A., Luna, B., Sweeney, J. A., and Eddy, W. F. (2002). Combining brains: a survey of methods for statistical pooling of information. *Neuroimage* **16**, 2 (Jun), 538–550. http://dx.doi.org/10.1006/nimg.2002.1107.

[21] Lettre, G. and Rioux, J. D. (2008). Autoimmune diseases: insights from genome-wide association studies. *Hum Mol Genet* **17**, R2 (Oct), R116–R121. http://dx.doi.org/10.1093/hmg/ddn246.

[22] Li, H., Wei, Z., and Maris, J. (2010). A hidden markov random field model for genome-wide association studies. *Biostatistics 11*, 139–150. http://dx.doi.org/10.1093/biostatistics/kxp043.

[23] Li, W., Meyer, C. A., and Liu, X. S. (2005). A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics 21 Suppl 1*, i274–i282. http://dx.doi.org/10.1093/bioinformatics/bti1046.

[24] Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal.* **47**, 3, 467–485. MR2086483

[25] Newton, M. A., Quintana, F. A., den Boon, J. A., Sengupta, S., and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.* **1**, 1, 85–106. http://dx.doi.org/10.1214/07-AOAS104. MR2393842

[26] Owen, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc., Ser. B, Stat. Methodol.* **67**, 3, 411–426. MR2155346

[27] Pyne, S., Futcher, B., and Skiena, S. (2006). Meta-analysis based on control of false discovery rate: combining yeast chip-chip datasets. *Bioinformatics* **22**, 20 (Oct), 2516–2522. http://dx.doi.org/10.1093/bioinformatics/btl439.

[28] Qiu, X., Klebanov, L., and Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Stat Appl Genet Mol Biol 4*, Article34. http://dx.doi.org/10.2202/1544-6115.1157. MR2183944

[29] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*. 257–286.

[30] Sabatti, C., Service, S., and Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics* **164**, 2 (Jun), 829–833.

[31] Sarkar, S. K. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *Ann. Stat.* **34**, 1, 394–415. MR2275247

[32] Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102**, 479, 901–912. MR2411657

[33] Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. *Journal Of The Royal Statistical Society Series B* **71**, 2, 393–424.

[34] Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S. F., Payne, F., Lowe, C. E., Szeszko, J. S., Hafler, J. P., Zeitels, L., Yang, J. H. M., Vella, A., Nutland, S., Stevens, H. E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L. J., Healy, B., Burren, O. S., Lam, A. A. C., Ovington, N. R., Allen, J., Adlem, E., Leung, H.-T., Wallace, C., Howson, J. M. M., Guja, C., Ionescu-Trgoviste, C., of Type 1 Diabetes in Finland, G., Simmonds, M. J., Heward, J. M., Gough, S. C. L., Consortium, W. T. C. C., Dunger, D. B., Wicker, L. S., and Clayton, D. G. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* **39**, 7 (Jul), 857–864. http://dx.doi.org/10.1038/ng2068.

[35] Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *American journal of human genetics* **81**, 6 (October), 1278–1283. http://dx.doi.org/10.1086/522374.

[36] Wei, Z. and Li, H. (2007). Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics* **8**, 2 (Apr), 265–284. http://dx.doi.org/10.1093/biostatistics/kxl007.

[37] Wei, Z., Sun, W., Wang, K., and Hakonarson, H. (2009). Multiple testing in genome-wide association studies via hidden markov models. *Bioinformatics* **25**, 21 (Nov), 2802–2808. http://dx.doi.org/10.1093/bioinformatics/btp476.

[38] Wu, W. B. (2008). On false discovery control under dependence. *Ann. Stat.* **36**, 1, 364–380. MR2387975

[39] Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method for combining p-values. *Genet Epidemiol* **22**, 2 (Feb), 170–185. http://dx.doi.org/10.1002/gepi.0042.

[40] Zhang, K., Qin, Z., Chen, T., Liu, J. S., Waterman, M. S., and Sun, F. (2005). Hapblock: haplotype block partitioning and tag snp selection software using a set of dynamic programming algorithms. *Bioinformatics* **21**, 1 (Jan), 131–134. http://dx.doi.org/10.1093/bioinformatics/bth482.

[41] Zhang, K., Sun, F., and Zhao, H. (2005). Haplore: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics* **21**, 1 (Jan), 90–103. http://dx.doi.org/10.1093/bioinformatics/bth388.

Wei Wang
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102, USA

Zhi Wei
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102, USA
E-mail address: zhiwei@njit.edu

Wenguang Sun
Department of Statistics
North Carolina State University
Raleigh, NC 27695, USA
E-mail address: sun@stat.ncsu.edu