Multiple testing in genome-wide association studies via hidden Markov models

Zhi Wei1,*, Wenguang Sun2, Kai Wang3 and Hakon Hakonarson3,4

1Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, 2Department of Statistics, North Carolina State University, Raleigh, NC 27695, 3Center for Applied Genomics, The Children’s Hospital of Philadelphia and 4Division of Genetics, Department of Pediatrics, The Children’s Hospital of Philadelphia, PA 19104, USA

Received on June 6, 2009; revised on July 17, 2009; accepted on July 31, 2009

ABSTRACT

Motivation: Genome-wide association studies (GWAS) interrogate common genetic variation across the entire human genome in an unbiased manner and hold promise in identifying genetic variants with moderate or weak effect sizes. However, conventional testing procedures, which are mostly P-value based, ignore the dependency and therefore suffer from loss of efficiency. The goal of this article is to exploit the dependency information among adjacent single nucleotide polymorphisms (SNPs) to improve the screening efficiency in GWAS.

Results: We propose to model the linear block dependency in the SNP data using hidden Markov models (HMMs). A compound decision-theoretic framework for testing HMM-dependent hypotheses is developed. We propose a powerful data-driven procedure (pooled local index of significance (PLIS)) that controls the false discovery rate (FDR) at the nominal level. PLIS is shown to be optimal in the sense that it has the smallest false negative rate (FNR) among all valid FDR procedures. By re-ranking significance with the Benjamini–Hochberg (BH) procedure, PLIS yields more accurate results and has better reproducibility of findings.

Conclusion: The genomic rankings based on our procedure are substantially different from the rankings based on the P-values. By integrating information from adjacent locations, the PLIS rankings benefit from the increased signal-to-noise ratio, hence our procedure often has higher statistical power and better reproducibility. It provides a promising direction in large-scale GWAS.

Availability: An R package PLIS has been developed to implement the PLIS procedure. Source codes are available upon request and will be available on CRAN (http://cran.r-project.org/).

Contact: zhiwei@njit.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Identifying genetic factors of complex diseases such as type 1 diabetes (T1D) can provide significant insights into the development of diagnostics and therapeutics of these diseases. For complex diseases, in addition to the hypothesized existence of a few genetic variants with strong effect, it is generally believed that there exist common moderate-risk variants that contribute interactively with significant influence. For example, it has been widely accepted that T1D disease is associated with a small number of genes with large effects and a large number of genes with small effects (Todd et al., 2007). Because of the recent advancements in comprehensive genomic information and cost-effective genotyping technologies, genome-wide association studies (GWAS) have become a popular tool to detect genetic variants that contribute to complex diseases. In recent years, GWAS have successfully identified 53 new common genomic regions that are associated with autoimmune diseases (Lette and Fioux, 2008), and have become a predominant tool as a first step to localize the unknown weak variants of complex diseases. However, the low statistical power has limited the practical advantage of GWAS. For example, the established genetic associations with T1D only explain part of the genetic risk for T1D. Many more genes with small to moderate effects remain to be discovered (Hakonarson et al., 2007; Todd et al., 2007).

In GWAS, it is typical to test hundreds of thousands of markers simultaneously. The family-wise error rate (FWER), defined as the probability of making at least one false rejection, is traditionally used as a measure to control the multiplicity. A well-known FWER procedure is the Bonferroni method. However, the power of a FWER controlling procedure is greatly reduced as the number of tests increases. In GWAS, where both the number of hypotheses and the number of true non-nulls may be large, it is cost-effective to tolerate some type I errors, provided that their number is small compared with the total number of discoveries. These considerations lead to a more powerful approach which calls for controlling the false discovery rate (FDR, Benjamini and Hochberg, 1995). The FDR, defined as the expected proportion of false rejections among all rejections, is a different view of how the errors in multiple comparisons can be combined. The FDR controlling procedures have been successfully applied in many large-scale studies such as multi-stage clinical trials, microarray experiments, brain imaging...
The works of Nyholt (2004) and Conneely and Boehnke (2007) showed that by exploiting the dependency structure, more precise results show that our procedure enjoys superior performance and yields the most accurate results in comparison with conventional approaches.

The rest of this article is organized as follows. Section 2 begins with an introduction of the HMM for SNP data, then introduces two approaches to testing groups of dependent hypotheses. Section 3 conducts simulation studies to compare the numerical performances of our approach versus conventional methods. In Section 4, our procedure is applied to the analysis of data from a GWAS of T1D for identifying disease-associated SNPs. We conclude the article with a discussion of results and methods. The proofs are given in the appendix in the Supplementary Material.

2 STATISTICAL METHODS

The HMM is an effective model to characterize the dependency among neighboring SNPs. In an HMM, each SNP has two hidden states: disease- or non-disease-associated, and the states of all SNPs along a chromosome form a Markov chain. The observed genotype data are generated conditionally on the hidden states via an observation model. This section first introduces some notation, then reviews an optimal FDR procedure by Sun and Cai (2009) for analysis of a single chromosome. Finally, two approaches to multiple-chromosome analysis are discussed.
2.1 A HMM for genotype data

Suppose there are \( n_0 \) cases and \( n_1 \) controls being genotyped over the \( m \) SNPs on chromosome \( k \). The total number of SNPs is \( n = \sum_k n_k \).

It is typical to first conduct a \( \chi^2 \)-test for each SNP to assess the association between the allele frequencies and the disease status, then obtain \( \alpha \)-values and \( z \)-values using appropriate transformations for further analysis. Let \( \theta_k = (\theta_{k,1}, \ldots, \theta_{k,m_k}) \) be the underlying states of the SNP sequence in chromosome \( k \) from the \( 5' \) to \( 3' \) end, where \( \theta_{k,m} = 1 \) indicates that SNP \( i \) in chromosome \( k \) is disease associated and \( \theta_{k,m} = 0 \) otherwise. We assume that \( \theta_k \) is distributed as a stationary Markov chain

\[
\theta_{k,j+1} = \frac{1}{\pi_{k,j}} \left[ 1 + \sum_{m=1}^{m_k} \text{LIS}(\theta_{k,m}) \right]^{\alpha_{k,j}},
\]

for \( k = 1, \ldots, K \). Let \( \pi_{k,j} = (1-\theta_{k,j})P_{k,0,1} + \theta_{k,j}P_{k,1,0} \). Denote by \( \mathbf{A}_k = (\omega_{k,j}) \) the transition matrix, \( \mathbf{A}_k = (\omega_{k,j,m}) \), the transition distribution; \( \mathbf{F}_k = (F_{k,0,1}, F_{k,1,0}) \), the observation distribution; and \( \mathbf{Y}_k = (\mathbf{A}_k, \mathbf{F}_k, \pi_k) \), the collection of all HMM parameters. Let \( \mathbf{Z}_k = (Z_{k,1}, \ldots, Z_{k,m_k}) \) be the observed \( z \)-values in chromosome \( k \) and \( \mathbf{E}_k = (\mathbf{Z}_k, \mathbf{F}_k, \pi_k) \). We assume that for a non-associated SNP, the \( z \)-value distribution is standard normal \( \mathcal{N}(0,1) \), and for a disease-associated SNP, the \( z \)-value distribution is normal mixture \( \mathcal{N}(k,\sigma_k^2) \). The normal mixture model can approximate a large collection of distributions and has been widely used (Magder and Zeger, 1996; Pan et al., 2004). When the number of components in the normal mixture \( k = 1 \) is known, the maximum likelihood estimate (MLE) of the HMM parameters can be obtained using the EM algorithm (Efron and Merhav, 2002; Sun and Cai, 2009). When \( k \) is unknown, we use the Bayesian information criterion (BIC) to select an appropriate \( k \).

2.2 Optimal FDR analysis in a single chromosome

This section introduces the optimal testing procedure for analysis of a single chromosome (hence \( k \) is suppressed in this subsection). Sun and Cai (2009) developed a compound decision-theoretic framework for testing HMM-dependent hypotheses and showed that the optimal testing procedure is a thresholding rule based on the LIS. The LIS, defined as \( L_k(i) = \hat{P}_k(i)/\hat{Q}_k(i,0) \), is the probability that a case is a null given the observed data. Let \( \mathbf{W} \) be an estimate of the HMM parameters. Denote by LIS(1), LIS(2),... LIS(\( \hat{K} \)), the ordered LIS values and \( H_{(1)}, \ldots, H_{(\hat{K})} \), the corresponding hypotheses. The LIS procedure operates as follows:

Let \( l = \max \left\{ \frac{1}{\hat{\alpha}} \sum_{k=1}^{\hat{K}} L_k(i) : 0 \right\} \) then reject all \( H_{(i)} \),

for \( i = 1, \ldots, \hat{K} \). Theorems 5 and 6 in Sun and Cai (2009) showed that under some regularity conditions, the LIS procedure is optimal in the sense that it controls the FDR at level \( \alpha \) and has the smallest FNR among all valid FDR procedures.

In practice, it is desirable to combine the testing results from several chromosomes so that the global (or genome-wide) FDR is controlled at the nominal level. Sections 2.3 and 2.4, respectively, consider a separate approach and a pooled approach that extend the LIS procedure for multiple-chromosome analysis.

2.3 Separate analysis of multiple chromosomes

The separate analysis (Efron, 2008) first conducts separately the analysis within each chromosome at the same FDR level \( \alpha \) and then combines the testing results from individual analyses. Define \( L_k(i) = P_k(i)/\hat{Q}_k(i,0) \). Denote by LIS(1), LIS(2),... LIS(\( \hat{K} \)), the ordered LIS values in chromosome \( k \) and \( H_{(1)}, \ldots, H_{(\hat{K})} \), the corresponding hypotheses. Inspired by the results in Section 2.2, it is natural to consider the following separated LIS (SLIS) procedure. For chromosome \( k \),

\[
\hat{e}_k = \max \left\{ \frac{1}{\hat{\alpha}} \sum_{i=1}^{m_k} L_k(i) : 0 \right\}.
\]

then reject all \( H_{(i)} \), \( i = 1, \ldots, \hat{K} \).

The final rejection set of the SLIS procedure, \( R_{\text{SLIS}} \), is obtained by combining the \( K \) rejection sets from all separate analyses: \( R_{\text{SLIS}} = \bigcup_{k=1}^{K} R_{(k)} \). The next theorem shows that the SLIS procedure is valid for FDR control.

**Theorem 1.** Consider the HMM defined by (1–3). Let \( L_k(i) \) be the ranked LIS values in chromosome \( k \), \( k = 1, \ldots, K \). Then the SLIS procedure controls the global FDR at level \( \alpha \).

2.4 Pooled analysis of multiple chromosomes

Although SLIS is valid for FDR control, it is inefficient in reducing the global FNR when the dependences in separate chromosomes are heterogeneous. A more powerful approach is the pooled LIS procedure (PLIS) derived in the appendix in the Supplementary Material. Under a compound decision-theoretic framework, we show that PLIS is asymptotically optimal.

Simulation studies conducted in Section 3 demonstrate that the performance of the 'do nothing' approaches, as well as the SLIS procedure, can be greatly improved by the PLIS procedure. The PLIS procedure operates in three steps:

Step 1: calculate the plug-in LIS statistic \( L_k(i) = P_k(i)/\hat{Q}_k(i,0) \) for individual chromosomes.

Step 2: combine and rank the plug-in LIS statistic from all chromosomes. Denote by \( L_{(1)}, \ldots, L_{(\hat{K})} \), the ordered values and \( H_{(1)}, \ldots, H_{(\hat{K})} \) the corresponding hypotheses.

Step 3: Reject all \( H_{(i)} \), \( i = 1, \ldots, \hat{K} \), where \( l \) is the smallest \( l \) such that \( \frac{1}{\hat{\alpha}} \sum_{i=1}^{\hat{K}} L_k(i) \leq l \).

**Remark 1.** PLIS is a hybrid strategy that has combined features from both pooled and separate analyses. Specifically, PLIS is a 'separate' analysis because, in Step 1, the grouping information is exploited to calculate chromosome-wise HMM parameters; PLIS 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 testing results from different chromosomes may be combined. Efron suggests using identical FDR levels for all chromosomes, whereas we suggest using different FDR levels, which are automatically adapted to the features of all groups.

The next theorem shows that PLIS is valid and asymptotically optimal. The proof of the theorem is outlined in the appendix in the Supplementary Material.

**Theorem 2.** Consider the HMM defined by (1–3). Let \( L_{(k)} \) be the ranked LIS values in chromosome \( k \), \( k = 1, \ldots, K \). Then the PLIS procedure controls the global FDR at level \( \alpha \). In addition, the global FNR of PLIS is \( \hat{\beta} + o(1) \), where \( \beta \) is the smallest FNR among all valid FDR procedures at level \( \alpha \).

Unlike for the separate analysis, the chromosome-wise FDR levels of the PLIS procedure are, in general, different from \( \alpha \). The actual chromosome-wise FDR levels of the PLIS procedure can be consistently estimated as \( \hat{\alpha}_k = (1/\hat{K}) \sum_{i=1}^{\hat{K}} \text{LIS}(\hat{K}) \), where \( \beta_k \) is the number of rejections in chromosome \( k \) and \( \text{LIS}(\hat{K}) \) is the \( i \)-th smallest statistic in chromosome \( k \). It is important to note that PLIS not only gives the optimal rankings of all hypotheses, but also suggests an optimal way of combining testing results from different chromosomes.

When \( \hat{\beta} \) is known, the LIS statistic can be computed in \( O(m) \) time [total \( O(\hat{K}) \) for the whole genome] using the forward-backward procedure (Rabiner, 1989). When \( \hat{\beta} \) is unknown, we first estimate the unknown quantities by \( \hat{\beta} \) using the EM algorithm, then plug-in \( \hat{\beta} \) to obtain \( \text{LIS}(\hat{K}) \).
To illustrate the main ideas and simplify discussion, we generated with transition matrices $\theta \sim \text{PLIS}$. Two Markov chains ($\theta$) dominated by PLIS. controlled, chromosome-wise FDRs may vary, inflated or deflated. ($\text{asymptotically, and BH is conservative; for PLIS, while its global FDR is}$ dominated by PLIS).

3 SIMULATION STUDIES

In this section, we investigate the numerical performance of the PLIS procedure and compare it with the SLIS procedure and the BH procedure (Benjamini and Hochberg, 1995). We will show that SLIS improves BH by exploiting the dependency information among adjacent SNPs, and PLIS further improves SLIS by optimally weighting the FDR levels among different chromosomes to minimize the overall FNR level.

3.1 Simulation I

To illustrate the main ideas and simplify discussion, we consider $K=2$ chromosomes, each of which has $m_1=m_2=2000$ SNPs. Two Markov chains ($\theta_1$) are first generated with transition matrices $\theta_1 = (0.98, 0.02; 0.3, 0.7)$ and $\theta_2 = (0.98, 0.02; 0.05, 0.95)$. Then the observations $(x_{k1})_{2000} \sim N(\mu_1, \theta_1)$ are generated conditional on $(\theta_1)$ namely, $x_{ki} \sim N(\mu_1, \theta_1)$. We vary $\mu_1$ from 1 to 4 with an increment 0.5 and let $\mu_2 = \mu_1 + 1$. Our goal is to find disease-associated SNPs, while controlling the FDR at a prespecified level for the whole genome (combining chromosomes 1 and 2). We apply the BH, SLIS and PLIS procedures at FDR level $\alpha = 0.1$. The simulation is repeated for 200 times.

The simulation results are summarized in Figure 1. From Figure 1a, we can see that the FDR levels of all three procedures are controlled at 0.10 asymptotically, and BH is conservative. We also plot together the FDR levels of PLIS for each individual chromosome. It is interesting to note that the FDR level for chromosome 1 is much greater than that of chromosome 2. This is because PLIS aims to minimize the genome-wise FNR while controlling the genome-wise FDR; the heterogeneity of individual chromosomes may lead to automatically adjusted chromosome-wise FDRs, inflated or deflated, in order to optimize the screening process. From Figure 1b, we can see that the FNR of BH is much higher than those of PLIS and SLIS, indicating that utilizing genomic dependency information can greatly improve the efficiency of a testing procedure. In addition, SLIS is dominated by PLIS, indicating that it is beneficial to use different FDR levels for different chromosomes. The difference in FNR levels (BH versus PLIS) becomes larger as $\mu$ decreases; this implies that PLIS is especially useful when the signals are weak.

Fig. 1. (a) The FDR levels of all three procedures are controlled at 0.10 asymptotically, and BH is conservative; for PLIS, while its global FDR is controlled, chromosome-wise FDRs may vary, inflated or deflated. (b) The FNR of BH is much higher than those of PLIS and SLIS, and SLIS is further dominated by PLIS.

Fig. 2. Ranking efficiency: PLIS has higher sensitivity than BH at the same FDR level; the improvement is more dramatic when signals are weak ($\mu$ is small).

3.2 Simulation II

In order to obtain more realistic linkage disequilibrium (LD) patterns among SNPs, we simulate a hypothetical dataset by sampling from the individuals in a case–control GWAS study conducted at the Children’s Hospital of Philadelphia (Hakonarson et al., 2007). We consider two regions, each having 2000 SNPs, from two different chromosomes. Two SNPs are selected from each region as the
K percentages of true positives (sensitivity) selected by the top as the three adjacent SNPs on each side of a causal SNP. From further analysis. The performance of a testing procedure is 1000 controls. The four-disease causal variants are then removed disease rate of 0.03. Based on this model, we repeatedly generate β where

\[ \beta = \frac{\log(1 + \frac{\alpha}{1 - \alpha})}{a} \]

causal SNPs are far away, while in the other region, the two causal disease causal SNPs (with relative risk 1.5). In one region, the two SNPs are close (separated by three SNPs). The disease probability can be described by using a logistic regression model:

\[ \Pr(Y = 1 | X) = \frac{\exp(\beta_0 + \sum_{i=1}^{4} \beta_i X_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{4} \beta_i X_i)} \]

where \( \beta_k = \log(1.5) \) for \( k = 1, \ldots, 4 \). We set \( \beta_0 = -0.48 \) to yield a disease rate of 0.03. Based on this model, we repeatedly generate the disease status for each individual until we obtain 1000 cases and 1000 controls. The four-disease causal variants are then removed from further analysis. The performance of a testing procedure is assessed by the selection rate of relevant SNPs, which are defined as the three adjacent SNPs on each side of a causal SNP.

We repeated the experiment 200 times and calculated the percentages of true positives (sensitivity) selected by the top K SNPs. The average sensitivity curves for comparing PLIS versus P-value are shown in Figure 3. We can see that the rankings by P-value are dominated by the rankings by PLIS.

In summary, the PLIS procedure, which exploits the HMM dependency, efficiently increases the signal-to-noise ratio in the sample by integrating information from adjacent locations. The precision is greatly improved in the sense that (i) the number of false positives is greatly reduced and (ii) 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 incorporated properly in a testing procedure.

4 APPLICATION TO T1D DATASETS

T1D is a highly heritable disease with heritability estimate of 90%. Compared with the general population, the first-degree relatives of patients with T1D are at 15-fold higher risk. It is hypothesized that multiple genetic loci contribute to the risk of developing T1D. However, the established genetic associations with T1D explain only little more than half of the genetic risk for T1D, indicating that the effects of other loci exist. To search systematically for these unknown loci, Hakonarson et al. (2007) performed a GWAS, where a discovery cohort, including 563 cases and 1146 controls, was collected. All participants were of European ancestry and recruited through pediatric diabetes clinics in Philadelphia, Montreal, Toronto, Ottawa and Winnipeg. The study subjects were genotyped using the Illumina HumanHap550 BeadChip at the Children’s Hospital of Philadelphia (CHOP). A replication cohort, consisting of 483 parents–offspring trios with affected children, was also collected and genotyped (Hakonarson et al., 2007; Grant et al., 2008).

A series of standard quality control procedures was performed to eliminate markers with minor allele frequency <1%, with Hardy–Weinberg equilibrium P < 1e-6, or with genotype no-call rate >5%. After the quality control, 534 213 markers on 23 chromosomes in the discovery cohort are eligible for further analysis. The same quality control procedure was applied to the replication cohort, but markers showing excessive (>4 families) Mendelian inconsistencies were eliminated. After the screening, 532 662 markers over 23 chromosomes in the replication cohort are eligible for further analysis. A detailed procedure for the analysis of GWAS datasets using PLIS is provided in the Supplementary Material.

4.1 Model selection and the estimation of HMM parameters

We first conducted a \( \chi^2 \)-test for each SNP to assess the association between the allele frequencies and the disease status, then obtain P-values and Z-values using appropriate transformations. Each chromosome is modeled separately to obtain chromosome-specific HMM parameters \( \Psi_k \) and LIS values \( \hat{L}_{k} \). We assume that the null distribution is standard normal \( N(0,1) \) and the non-null distribution is a normal mixture \( \sum_{i=1}^{m} \alpha_i N(0, \sigma_i^2) \). The number of components \( L \) in the non-null distribution is determined by the BIC criterion, \( \text{BIC} = \log(P(\hat{\Psi}_k | \xi_k)) - |\hat{\Psi}_k| \log(m) \), where \( P(\hat{\Psi}_k | \xi_k) \) is the likelihood function, \( \hat{\Psi}_k \) is the MLE of HMM parameters and \( |\hat{\Psi}_k| \) is the number of HMM parameters. We vary \( L \) and evaluate different choices of \( L \) for each chromosome. The high transition probabilities \( \left( \alpha_{00}, \alpha_{11} \right) \) are observed, which suggests strong genomic dependency. The wide range of \( \alpha_{11} \) from 0.89 to 0.97 shows that the chromosomal dependencies are heterogeneous, which
The improvement is quite substantial when... eight are identified by PLIS. Detailed results... identified by BH and eight are identified by PLIS. Detailed results... the corresponding... SNPs from the HLA region. A total of 1841 SNPs, including 419 SNPs from the human leukocyte antigen (HLA) region, are... the LIS rankings benefit... The reproducibility is defined as the proportion of top \( k \) SNPs in the discovery dataset that are also on the top-\( k \) list of the replication dataset. Justifies the use of chromosome-specific HMMs. In addition, we find a two- or three-component normal mixture model that is sufficient in most situations. The complete estimated HMM parameters with corresponding optimal choice of \( L \) are summarized in Table 1 in the Supplementary Material.

4.2 Superiority in ranking SNPs

GWAS is often criticized for its poor reproducibility. It is common that, for the same disease, a large proportion of SNPs claimed to be significant in one GWAS are not significant in another GWAS. The genomic rankings based on the LIS statistics are substantially different from the rankings based on \( P \)-values. By integrating information from adjacent locations, the LIS rankings benefit from the increased signal-to-noise ratio, hence often has better reproducibility. To illustrate this, we first pool the LIS statistics across all chromosomes together and rank them in an increasing order. The reproducibility is defined as the proportion of top \( k \) SNPs in the discovery dataset that are also on the top-\( k \) list of the replication dataset. Then, we vary \( k \) from 50 to 5000 and plot the reproducibilities of the LIS statistics and \( P \)-values as functions of \( k \). The results are summarized in Figure 4. We can see that the reproducibility of LIS rankings is better for all choices of \( k \). The improvement is quite substantial when \( k \) is several hundreds, where the corresponding \( P \)-values range from \( 10^{-3} \) to \( 10^{-7} \). This magnitude of \( P \)-value often corresponds to small to moderate effects that remain to be discovered/confirmed.

4.3 Results and discussion

To quantify the overall evidence for association, we combine the \( P \)-values from the discovery cohort and the replication cohort using Fisher’s method (Fisher, 1958). The 531,026 combined \( P \)-values were then transformed to \( Z \)-values. We apply both the BH and PLIS procedures at FDR-level 0.0001. A total of 1841 SNPs, including 1217 SNPs from the human leukocyte antigen (HLA) region, are identified by PLIS as disease associated. Most of these SNPs appear in clusters. As a comparison, the BH procedure identifies 419 SNPs at the same FDR level, and 387 of them from the HLA region. A meta-analysis based on recent GWAS has confirmed 46 T1D susceptibility loci (Barrett et al., 2007). In contrast, PLIS does not claim two loci, namely the gene for collagen type 1 \( \alpha2 \) (COL1A2; rs10255021) and rs672797, in the vicinity of latrophilin 2 (LPHN2), are disease-associated, while the \( P \)-value-based approaches even claimed their significant association under the Bonferroni correction (Hakonarson et al., 2007). However, these two loci failed to replicate in follow-up studies. A closer look at the nearby regions shows that these two SNPs are both surrounded by insignificant SNPs; hence the ‘significance’ is more likely to be ‘noise’. By borrowing information from nearby locations, PLIS successfully classify them as non-disease-associated SNPs.

5 CONCLUSION AND DISCUSSION

This article develops HMMs for analysis of the SNP data arising from large-scale GWAS. The LIS procedure in Sun and Cai (2009) is extended for multiple-chromosome analysis. We show that by incorporating the HMM dependency, the accuracy of a multiple testing procedure can be greatly improved. The numerical performances of our GLIS procedure are investigated using both simulated and real data. It is demonstrated that our PLIS procedure, compared with conventional ‘do nothing’ approaches, is more powerful in identifying small to moderate effects and has better reproducibility of results.

The PLIS procedure can be improved in several ways. First, it is known that the SNP dependency does not decrease monotonically with physical distance. In general, a network would be a more precise description of the complex SNP dependency. It would be of interest to generalize our testing procedure from a Markov chain to a Markov random field. Second, instead of summarizing the data for each SNP using a \( z \)-value, we may consider a parametric hierarchical multinomial-Dirichlet model. Such hierarchical Bayes models allow a level of information sharing amongst the SNPs, which often lead to better inference procedures. Finally, it might be a strong assumption that the whole chromosome follow a stationary Markov chain. When applying PLIS, the practitioner can further divide a chromosome into smaller regions based on their domain...
knowledge, e.g. LD block structure derived from HapMap. Then the pooling concept in PLIS can be applied to different regions, which are assumed to follow region-specific HMMs. We expect that PLIS can be further improved by working on finer scales and more homogeneous regions.

Conflict of Interest: none declared.

REFERENCES


