Multiple Testing in Genome-Wide Association Studies via Hidden Markov Models
Zhi Wei 1*, Wenguang Sun 2, Kai Wang 3 and Hakon Hakonarson 3,4

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, Philadelphia, PA 19104
4Division of Genetics, Department of Pediatrics, The Children’s Hospital of Philadelphia, Philadelphia, PA 19104

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 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. A compound decision-theoretic framework for testing HMM-dependent hypotheses is developed. We propose a powerful data-driven procedure (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 for all SNPs with dependency considered, PLIS gains higher power than conventional p-value based methods. Simulation results demonstrate that PLIS dominates conventional FDR procedures in detecting disease associated SNPs. Our method is applied to analysis of the SNP data from a GWAS of type 1 diabetes. Compared to the BH procedure, PLIS yields more accurate results and has better reproducibility of findings.

Conclusion: The genomic rankings based on the 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. This provide 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

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 (Lettre and Froux, 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 (Todd et al., 2007 and Hakonarson 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 to 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 studies and astronomical surveys, among others (Efron et al., 2001; Miller et al., 2001; Tusher et al., 2001; Storey and Tibshirani, 2003; Dudoit et al., 2002; Sabatti et al., 2003; Meinshausen and Rice,
procedures are based on thresholding the ranked \( p \)-values (Benjamini and Hochberg, 1995; Storey, 2002; Benjamini and Hochberg, 2000; Genovese and Wasserman, 2002; van der Laan et al., 2004). Conventional \( p \)-value-based procedures are not successful in identifying “weak” genes from large-scale GWAS data: a liberal \( p \)-value cutoff yields too many false positives, yet a strict cutoff tends to wipe out most moderate effects. Many useful features in the SNP data can actually be exploited to increase power but have been largely ignored by conventional \( p \)-value methods. One important feature is the linear block structure among the SNPs. Specifically, the crossover during meiosis will break parental chromosomes into inheritable segments and then form gametes by recombining those segments. As a result, adjacent genomic loci tend to co-segregate and it is common to observe that disease-associated SNPs form clusters and exhibit high correlation.

The correlation effects on FDR analyses have been extensively studied (Owen, 2005; Sarkar, 2006; Qiu et al., 2005; Efron, 2007). We call an FDR procedure is valid if it controls the FDR at the nominal level and optimal if it has the smallest false negative rate (FNR). Genovese and Wasserman (2002) among all valid FDR procedures. Here the FNR, defined as the expected proportion of falsely accepted hypotheses, is a similar concept as the type II error rate in single hypothesis testing. Most previous research has been focused on the validity issue. For example, it was shown that the BH procedure is valid under different dependence assumptions (Benjamini and Yekutieli, 2001; Farcomeni, 2007; Wu, 2009). Sabatti et al. (2003) observed that the validity holds for the BH procedure in case-control studies. These methods, referred to as the “do nothing” approaches, ignore the dependency information and treat all dependent tests as if they were independent. The “do nothing” approaches, although valid, may suffer from severe loss of efficiency. For example, Sabatti et al. (2003) found that the BH procedure suffers from increased power loss with increased dependency among markers in a genome scan. The works of Nyholt et al. (2004) and Conneely et al. (2007) showed that by exploiting the dependency structure, more precise FDR control can be achieved and hence the statistical power can be improved.

The development of a multiple testing procedure essentially involves two steps: ranking the hypotheses and choosing a cutoff along the rankings. The ranking step is more fundamental. Different from Nyholt et al. (2004) and Conneely et al. (2007) that utilize the dependency to choose the cutoff, we propose to utilize the dependency to create more efficient rankings. The proposed procedure uniformly improves all \( p \)-value based procedures by re-ranking the importance of all SNPs. It has been appreciated that genomic dependency information can significantly improve the efficiency in analysis of large-scale genomic data (Wei and Li, 2007; Wei and Li, 2008). We expect that the information of the SNP dependency can be exploited to construct more efficient tests. From a biological point of view, the SNP dependency is informative in constructing more efficient association tests because, when a SNP is associated with a disease, 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. In this article, the dependency of adjacent SNPs is captured using a hidden Markov Model (HMM).

The first goal of this article is to show that conventional “do nothing” procedures can be greatly improved by exploiting the HMM dependency. Related theoretical results for testing HMM-dependent hypotheses are developed in Sun and Cai (2009), in which an optimal procedure (LIS) is proposed. However, LIS only deals with a single Markov chain and cannot be applied directly to GWAS. The reason is that the chromosomes in a genome segregate independently and may exhibit different dependency structures. It is therefore more appropriate to model each chromosome separately. The second goal of this article is to extend Sun and Cai’s approach to deal with several groups of dependent tests and to develop a new testing procedure for multiple-chromosome analysis. The key issue is how to combine the simultaneous inferences made for separate chromosomes to achieve optimal genome-wise FDR control.

A straightforward approach to combining the analyses from different chromosomes is the so-called separate analysis (Efron 2008), which suggests applying the LIS procedure to each chromosome at the same FDR level. However, this approach is not optimal in general. In this article, we propose a pooled procedure (PLIS) for multiple-chromosome analysis. It is shown that PLIS is optimal in the sense that the genome-wise FNR is minimized subject to a constraint on the genome-wise FDR. One important feature of PLIS is that different chromosome-wise FDR levels are chosen such that genome-wise FNR level is minimized. The validity and optimality of PLIS are established in a compound decision-theoretic framework. The numerical performances of PLIS are investigated using both simulated and real data. The numerical 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 vs. 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 document.

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-associated 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 Hidden Markov Model for Genotype Data

Suppose there are \( n_1 \) cases and \( n_2 \) controls being genotyped over the \( n_1 \) SNPs on chromosome \( k, k = 1, \ldots, K \). The total number of SNPs is \( m = \sum_{k=1}^{K} m_k \). 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 \(p\)-values and \(z\)-values using appropriate transformations for further analysis.

Let \(\theta_k = \{\theta_{k1}, \cdots, \theta_{km_k}\}\) be the underlying states of the SNP sequence in chromosome \(k\) from the 5’ end to the 3’ end, where \(\theta_{ki} = 1\) indicates that SNP \(i\) from chromosome \(k\) is disease-associated and \(\theta_{ki} = 0\) otherwise. We assume that \(\theta_k\) is distributed as a stationary Markov chain

\[
P(\theta_k | \theta_{k-1}, \theta_{k-2}, \cdots) = \prod_{i=1}^{m_k} P(\theta_{ki} | \theta_{k,i-1})
\]

with transition probability \(a_{kij} = P(\theta_{ki} = j | \theta_{ki-1} = i)\), and that \(\theta_k\) and \(\theta_k\) are independent for \(i \neq j\).

i.e., SNPs on different chromosomes are independent. In an HMM, the observed data are assumed to be conditionally independent given the hidden states:

\[
P(\mathbf{z}_k | \theta_k, \mathbf{F}_k) = \prod_{i=1}^{m_k} P(\mathbf{z}_{ki} | \theta_{ki}, \mathbf{F}_k),
\]

for \(k = 1, \cdots, K\). Let \(Z_{ki} | \theta_{ki} \sim (1-\theta_{ki}) F_{k0} + \theta_{ki} F_{k1}\). Denote by \(A_k = (a_{kij})\) the transition matrix, \(\pi_k = (\pi_{k,0}, \pi_{k,1})\) the stationary distribution, \(F_k = (F_{k0}, F_{k1})\) the observation distribution, and \(\Psi_k = (A_k, \pi_k, F_k)\) the collection of all HMM parameters. Let \(\mathbf{z}_k = (z_{1k}, \cdots, z_{mk})\) be the observed \(z\)-values in chromosome \(k\) and \(\mathbf{z} = (z_1, \cdots, z_K)\). We assume that for a non-associated SNP, the \(z\)-value distribution is standard normal \(F_{k0} = N(0, 1)\), and for disease-associated SNPs the \(z\)-value distribution is a normal mixture \(F_{k1} = \sum_{m=1}^{M_k} N(\mu_m, \sigma_m^2)\). The normal mixture model can approximate a large collection of distributions and has been widely used (Magder and Zeger, 1996; Pan et al., 2003; Efron, 2004). When the number of components in the normal mixture \(L_k\) is known, the maximum likelihood estimate (MLE) of the HMM parameters can be obtained using the EM algorithm (Ephraim and Merhav, 2002; Sun and Cai, 2009). When \(L_k\) is unknown we use the Bayesian information criterion (BIC) to select an appropriate \(L_k\).

### 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 local index of significance (LIS). The LIS, defined as \(\text{LIS}_i = P(\theta_{ki} = 1 | \mathbf{z}_i)\), is the probability that a case is a null given the observed data. Let \(\Psi\) be an estimate of the HMM parameters. Denote by \(\text{LIS}_{(1)}, \cdots, \text{LIS}_{(m)}\) the ordered LIS values and \(H_{(1)}, \cdots, H_{(m)}\) the corresponding hypotheses. The LIS procedure operates as follows:

Let \(l = \max \{i : \sum_{j=1}^{i} \text{LIS}_{(j)} \leq \alpha\}\). Then reject all \(H_{(i)}\), for \(i = 1, \cdots, l\). Theorem 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. Subsections 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 \(\text{LIS}_{(i)} = P(\theta_{ki} = 1 | \mathbf{z}_i)\). Denote by \(\text{LIS}_{(1)}, \cdots, \text{LIS}_{(m)}\) the ordered LIS values in chromosome \(k\) and \(H_{(1)}, \cdots, H_{(m)}\) the corresponding hypotheses. Inspired by the results in Section 2.2, it is natural to consider the following separated LIS (SLIS) procedure. For chromosome \(k\),

\[
\text{let } l_k = \max \{i : \frac{1}{l} \sum_{j=1}^{i} \text{LIS}_{(j)} \leq \alpha\},
\]

then reject all \(H_{(i)}, i = 1, \cdots, l_k\).

The final rejection set of the SLIS procedure, \(\mathcal{R}_{\text{SLIS}}\), is obtained by combining the \(K\) rejection sets from all separate analyses:

\[
\mathcal{R}_{\text{SLIS}} = \bigcup_{k=1}^{K} \{H_{(i)} : i = 1, \cdots, l_k\}.
\]

The next theorem shows that the SLIS procedure is valid for FDR control.

**Theorem 1.** Consider the HMM defined by (1)-(3). Let \(\text{LIS}_{(i)}\) be the ranked LIS values in chromosome \(k, k = 1, \cdots, K\). 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 dependencies in separate chromosomes are heterogeneous. A more powerful approach is the pooled LIS procedure (PLIS) derived in the appendix. 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 \(\text{LIS}_{(i)} = P(\theta_{ki} = 1 | \mathbf{z}_i)\) for individual chromosomes.

**Step 2.** Combine and rank the plug-in LIS statistic from all chromosomes. Denote by \(\text{LIS}_{(1)}, \cdots, \text{LIS}_{(m)}\) the ordered values and \(H_{(1)}, \cdots, H_{(m)}\) the corresponding hypotheses.

**Step 3.** Reject all \(H_{(i)}, i = 1, \cdots, l\), where \(l = \max \{i : (1/l) \sum_{j=1}^{i} \text{LIS}_{(j)} \leq \alpha\}\).

**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.

**Theorem 2.** Consider the HMM defined by (1)-(3). Let \(\text{LIS}_{(i)}\) be the ranked LIS values in chromosome \(k, k = 1, \cdots, K\). Then
the PLIS procedure controls the global FDR at level $\alpha$. In addition, the global FNR level of PLIS is $\beta' + o(1)$, where $\beta'$ is the smallest FNR level 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 $\text{FDR} = (1/R_k) \sum_{i=1}^{R_k} \text{PLIS}_k(i)$, where $R_k$ is the number of rejections in chromosome $k$ and PLIS$_k(i)$ 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 $\Psi_k$ is known, the LIS statistic can be computed in $O(m_k)$ time (total $O(Km_k^2)$ for the whole genome) using the forward-backward procedure (Rabiner, 1989). When $\Psi_k$ is unknown, we first estimate the unknown quantities by $\hat{\Psi}_k$ using the EM algorithm, then plug-in $\hat{\Psi}_k$ to obtain LIS$_k$.

### 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 main ideas and simplify discussion, we consider $K = 2$ chromosomes, each of which has $m_1 = m_2 = 2000$ SNPs. Two Markov chains $(\theta_{k1}, \theta_{k2})_{i=1}^{2000}$, $k = 1, 2$, are firstly generated with transition matrices $A_1 = (0.98, 0.02; 0.3, 0.7)$ and $A_2 = (0.98, 0.02; 0.05, 0.95)$. Then the observations $(x_{k1}, x_{k2}, \ldots, x_{k1})_{i=1}^{2000}$ are generated conditional on $(\theta_{k1}, \theta_{k2})_i \sim N(0, 1)$ and $N(\mu_k, 1)$. 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 pre-specified level $\alpha$. Simulations are repeated for 200 times. The simulation results are summarized in Figure 1. From Panel (a), 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-wide FNR while chromosome 1 is much more informative than chromosome 2. To illustrate main ideas and simplify discussion, we consider $K = 2$ chromosomes, each of which has $m_1 = m_2 = 2000$ SNPs. Two Markov chains $(\theta_{k1}, \theta_{k2})_{i=1}^{2000}$, $k = 1, 2$, are firstly generated with transition matrices $A_1 = (0.98, 0.02; 0.3, 0.7)$ and $A_2 = (0.98, 0.02; 0.05, 0.95)$. Then the observations $(x_{k1}, x_{k2}, \ldots, x_{k1})_{i=1}^{2000}$ are generated conditional on $(\theta_{k1}, \theta_{k2})_i \sim N(0, 1)$ and $N(\mu_k, 1)$. 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 pre-specified level $\alpha$. Simulations are repeated for 200 times. The simulation results are summarized in Figure 1. From Panel (a), 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-wide FNR while controlling the genome-wide 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 Panel (b), 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 vs. PLIS) becomes larger as $\mu$ decreases; this implies that PLIS is especially useful when the signals are weak.

It is important to point out that the higher power of PLIS is not gained at the price of a higher FDR level; this can be illustrated by Figure 2. Here, the sensitivity is calculated as the average proportions of correctly identified SNPs over the 200 replications. We vary the significance thresholds and calculate corresponding FDRs and sensitivities. We can see that PLIS discovers more true disease-associated SNPs in clusters while the BH procedure can only identify sporadic suspicious SNPs.

### Table 1. The significance levels suggested by BH vs. PLIS

<table>
<thead>
<tr>
<th>Sequence</th>
<th>States</th>
<th>p-values</th>
<th>LIS Values</th>
<th>BH Procedure</th>
<th>PLIS Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>2694</td>
<td></td>
<td>3.62e-01</td>
<td>1.59e-03</td>
<td>$\bigcirc$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td>2695</td>
<td></td>
<td>1.52e-03</td>
<td>6.85e-05</td>
<td>$\bullet$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td>2696</td>
<td></td>
<td>6.62e-04</td>
<td>9.29e-04</td>
<td>$\bullet$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td>2697</td>
<td></td>
<td>7.19e-01</td>
<td>1.00e-01</td>
<td>$\bullet$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>2718</td>
<td></td>
<td>1.23e-02</td>
<td>1.65e-04</td>
<td>$\bigcirc$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td>2719</td>
<td></td>
<td>3.37e-03</td>
<td>8.19e-05</td>
<td>$\bullet$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td>2720</td>
<td></td>
<td>5.59e-01</td>
<td>2.19e-03</td>
<td>$\bullet$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td>2721</td>
<td></td>
<td>2.07e-04</td>
<td>2.67e-05</td>
<td>$\bigcirc$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td>2722</td>
<td></td>
<td>3.42e-02</td>
<td>7.42e-04</td>
<td>$\bigcirc$</td>
<td>$\bullet$</td>
</tr>
<tr>
<td>2723</td>
<td></td>
<td>2.88e-01</td>
<td>6.19e-03</td>
<td>$\bullet$</td>
<td>$\bullet$</td>
</tr>
</tbody>
</table>

$\bigcirc$ denotes a null hypothesis or an acceptance and $\bullet$, a non-null hypothesis or a rejection. By borrowing information from adjacent SNPs, PLIS tends to select disease-associated SNPs in clusters while the BH procedure can only identify sporadic suspicious SNPs.

The superiority of PLIS is achieved by incorporating the informative dependency structure and hence producing more...
efficient rankings of the SNPs. The PLIS rankings are fundamentally different from the rankings by BH. Table 1 compares the outcomes of the PLIS and BH procedures for testing two clusters of non-null SNPs. It is interesting to note that BH suggests site 2720 be less significant than site 2723. In contrast, PLIS suggests that site 2720, being surrounded by significant neighbors, be more significant than site 2723. By borrowing information from adjacent SNPs, the LIS statistic is more robust against local disturbance, which enables the PLIS procedure to select disease-associated SNPs in clusters. In contrast, the BH procedure can only identify sporadic suspicious SNPs.

3.2 Simulation II

In order to obtain more realistic LD patterns among SNPs, we simulate a hypothetical data set 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 disease causal SNPs (with relative risk 1.5). In one region, the two causal SNPs are far away, while in the other region, the two causal SNPs are close (separated by 3 SNPs). The disease probability can be described 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 $\beta_0 = -4.48$ 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 4 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 3 adjacent SNPs on each side of a causal SNP.

We repeated the experiment for 200 times and calculate the percentages of true positives (sensitivity) selected by the top $K$ SNPs. The average sensitivity curves for comparing PLIS versus BH 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 (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 incorporated properly in a testing procedure.

4 APPLICATION TO TYPE 1 DIABETES DATASETS

T1D is a highly heritable disease with heritability estimate of 90%. Compared to 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 genome-wide association study, where a discovery cohort, including 563 cases and 1146 controls, was collected. All participants were of European ancestry and recruited through paediatric 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 less than 1%, with Hardy-Weinberg Equilibrium $p$-values lower than 1e-6, or with genotype no-call rate higher than 5%. After the quality control, 534213 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
4.3 Results and Discussion

that remain to be discovered/confirmed. The magnitude of Fisher’s method (Fisher, 1958). The 531,026 combined p-values from the discovery cohort and the replication cohort using reproducibility of LIS rankings is better for all choices of k. The reproducibility is defined as the proportion of top k SNPs in the discovery list of the replication dataset.

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 LIS$_{k,i}$. We assume that the null distribution is standard normal $\mathcal{N}(0, 1)$ and the non-null distribution is a normal mixture $\sum_{l=1}^{L} c_l \mathcal{N}(\mu, \sigma^2_l)$. The number of components $L$ in the non-null distribution is determined by the BIC criterion, $\text{BIC} = \log \{P(\Psi_k | x)\} - \frac{1}{2} \log(m)$, where $P(\Psi_k | x)$ is the likelihood function, $\Psi_k$ is the MLE of HMM parameters, and $|\Psi_k|$ is the number of HMM parameters. We vary $L$ and evaluate different choices of $L$ for each chromosome. The high transition probabilities $a_{00}$ and $a_{11}$ are observed, which suggests strong genomic dependency. The wide range of $a_{11}$ (from 0.89 to 0.97) shows that the chromosomal dependencies are heterogeneous, which justifies the use of chromosome specific HMMs. In addition, we find a two or three-component normal mixture model is sufficient in most situations. The complete estimated HMM parameters with corresponding optimal choice of $L$ are summarized in Table 1 in the supplementary document.

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 Fig 4. We can see that the reproducibility of LIS rankings is better for all choices of k’s. The improvement is quite substantial when k is several hundreds, where the corresponding p-values range from $10^{-5}$ 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 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, 2009), among which three are identified by BH and eight are identified by PLIS. Detailed results are provided in Table 2. We observe significant improvement of rankings for T1D loci by PLIS over those by p-value. Such optimal rankings are very helpful for downstream analysis, such as the aforementioned meta-analysis and the recent popular gene/pathway based analysis (Wang et al, 2007). In contrast, PLIS does not claim two loci, namely the gene for collagen type 1 a2 (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 to 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
Multiple Testing via HMMs

Table 2. The 8 known T1D susceptibility loci identified by PLIS.

<table>
<thead>
<tr>
<th>Chr</th>
<th>T1D Loci</th>
<th>SNP</th>
<th>Dst</th>
<th>LIS Stat</th>
<th>LIS Rank</th>
<th>P Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>rs3129871*</td>
<td>rs3129871</td>
<td>0</td>
<td>2.4e-89</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>rs2476601</td>
<td>rs2476601</td>
<td>0</td>
<td>3.2e-09</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>rs725613</td>
<td>rs725613</td>
<td>2.7e-08</td>
<td>6</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>rs4244808</td>
<td>rs4320932 8</td>
<td>8.4e-05</td>
<td>128</td>
<td>956</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>rs3024505</td>
<td>rs1554286 4</td>
<td>1.6e-04</td>
<td>223</td>
<td>1398</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>rs3757247</td>
<td>rs10498965</td>
<td>42</td>
<td>2.0e-04</td>
<td>288</td>
<td>1350</td>
</tr>
<tr>
<td>6</td>
<td>rs9388489</td>
<td>rs1490388 137</td>
<td>2.3e-04</td>
<td>335</td>
<td>4359</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>rs2664170</td>
<td>rs884840 18</td>
<td>3.6e-04</td>
<td>608</td>
<td>7390</td>
<td></td>
</tr>
</tbody>
</table>

Three loci are identified directly; five loci are identified by nearby SNPs in complete LD with them (D'≥1, the disequilibrium score of the significant SNPs to the known T1D loci, derived from HapMap CEPH Utah data); Chr: chromosome, Dst: distance of the significant SNPs to the known T1D loci, in Kb; LIS Rank, P Rank: the ranks by LIS statistic and p-value, respectively, excluding the SNPs from the HLA region, *except for rs3129871, which locates in the HLA region.


Grant, S.F. et al. (2008), Follow up analysis of genome-wide association data identifies novel loci for type 1 diabetes, Diabetes, 58, 290–295.


