Xiao et al. BMC Bioinformatics 2013, 14:282
http://www.biomedcentral.com/1471-2105/14/282

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Large-scale multiple testing in genome-wide association studies via region-specific hidden Markov models

Jian Xiao, Wensheng Zhu* and Jianhua Guo

Abstract

Background: Identifying genetic variants associated with complex human diseases is a great challenge in genome-wide association studies (GWAS). Single nucleotide polymorphisms (SNPs) arising from genetic background are often dependent. The existing methods, i.e., local index of significance (LIS) and pooled local index of significance (PLIS), were both proposed for modeling SNP dependence and assumed that the whole chromosome follows a hidden Markov model (HMM). However, the fact that SNP data are often collected from separate heterogeneous regions of a single chromosome encourages different chromosomal regions to follow different HMMs. In this research, we developed a data-driven penalized criterion combined with a dynamic programming algorithm to find change points that divide the whole chromosome into more homogeneous regions. Furthermore, we extended PLIS to analyze the dependent tests obtained from multiple chromosomes with different regions for GWAS.

Results: The simulation results show that our new criterion can improve the performance of the model selection procedure and that our region-specific PLIS (RSPLIS) method is better than PLIS at detecting disease-associated SNPs when there are multiple change points along a chromosome. Our method has been used to analyze the Daly study, and compared with PLIS, RSPLIS yielded results that more accurately detected disease-associated SNPs.

Conclusions: The genomic rankings based on our method differ from the rankings based on PLIS. Specifically, for the detection of genetic variants with weak effect sizes, the RSPLIS method was able to rank them more efficiently and with greater power.

Background

At the present time, genome-wide association studies (GWAS) have become a very popular tool for identifying novel genetic variants for complex traits. GWAS typically tests hundreds of thousands of markers simultaneously, making it necessary to improve the power of large-scale multiple testing. Fortunately, the false discovery rate (FDR) for controlling such procedures, which was introduced in a seminal paper [1], is one of the most important methodological developments in multiple hypothesis testing and has played successful role in many large-scale multiple testing studies. Such studies include multi-stage clinical trials, microarray experiments, brain imaging studies, and astronomical surveys, amongst others [2-10]. Recently, the FDR approach has also been applied to GWAS [11]. Naturally, despite the increasing popularity of FDR, most of the traditional analytical methods for GWAS with FDR control have largely been proposed for individual single nucleotide polymorphism (SNP) analysis. However, because SNPs on the same chromosome are in local linkage disequilibrium (LD), which results in the complex dependence and correlation among large-scale tests, the traditional FDR controlling procedure for independent tests can potentially be conservative and lead to a loss of power. Therefore, it is important to consider FDR control for multiple testing procedures when the tests are dependent in GWAS.

Fortunately, Wei and Li pointed out that genomic dependency information could significantly improve the efficiency of analysis of large-scale genomic data [12,13]. We also expect that information about SNP dependency can be exploited to construct tests that are more efficient.

*Correspondence: wszhu@nenu.edu.cn
Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China

From a biological point of view, SNP dependency is informative when 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 (owing to the co-segregation). Therefore, when deciding the significance level of a SNP, its neighboring SNPs should be taken into account. Sun and Cai [14] proposed a local index of significance (LIS) controlling procedure that uses a hidden Markov model (HMM) to represent the dependence structure, and has shown its optimality under certain conditions and its strong empirical performance. Furthermore, this LIS procedure was extended to pooled local index of significance (PLIS) procedure for multiple-chromosome analysis [15], where the authors developed chromosome-specific HMMs for analysis of the SNP data arising from large-scale GWAS. Instead of HMM, Li [16] introduced a hidden Markov random field model (HMRFM) to account for LD when analyzing the SNP data from GWAS.

Therefore, the above methods are all based on the strong assumption that each chromosome follows a HMM or HMRFM. However, there are usually various LD patterns or haplotype blocks in the data, which result in heterogenity of dependencies among SNPs and variations in the disease risk rates of casual alleles in the different blocks. Hence, we suggest that the different blocks of each chromosome should follow different HMMs. Wei et al. [15] have stated that the development of a multiple testing procedure essentially involves two steps: ranking the hypotheses and choosing a cutoff along the rankings, where the ranking step is more fundamental. Obviously, modeling different regions by different HMMs can improve the efficiency of ranking. To this end, we should first identify change points for each chromosome, which can be used to divide the whole chromosome into smaller homogeneous regions. Specifically, we need to determine the number of change points as well as their locations on chromosomes.

In addition, the existing methods [14,15] assume that the observation variables follow a normal mixture distribution conditional on the latent variables in HMM, where the number of components is unknown in the normal mixture distribution. Sun and Cai [14] showed that the number of components should be determined by the likelihood-based Bayesian information criterion (BIC). However, BIC, as well as many other existing criterions, rely on a strong assumption that the observations are independent and require large sample sizes to reach their asymptotic consistency behavior. While in HMM, the observations are dependent such that the effective sample size for these criterions may be small.

In this paper, we first focus on the problem of how to infer simultaneously the number of components in a normal mixture distribution as well as the change points of each chromosome. We put forward a data-driven penalized criterion for model selection in HMM, and propose a sliding window-based improved version of the dimension jump method [17] to estimate this criterion. We then applied the dynamic programming (DP) algorithm to find multiple change points. The numerical results show that the proposed adaptive criterion has better performance than the original version. Second, we extended the approach of Wei et al. [15] to develop a testing procedure, which we have called region-specific PLIS (RSPLIS), for the analysis of different chromosomes with multiple regions. The numerical results show that RSPLIS outperforms PLIS in a disease association study. Our proposed procedure has been used to analyze the data from Daly et al. [18] for identifying Crohn's disease-associated SNPs.

Methods

First, Sun and Cai [14] developed a compound decision-theoretic framework for testing HMM-dependent hypotheses and presented an optimal testing procedure that can be used to analyze a single chromosome for SNP data. Second, Wei et al. [15] proposed a PLIS approach for multiple-chromosome analysis. They showed that under some regularity conditions, the PLIS procedure is valid and asymptotically optimal in the sense that it can control the global chromosome-wise FDR at the nominal level α and has the smallest false non-discovery rate (FNR) among all valid FDR procedures by combining the testing results from different chromosomes. In this section, we extend the chromosome-specific PLIS to the analysis of different chromosomes with multiple regions. In what following, we formulate our statistical model and elaborate the theoretical foundations of our method.

Region-specific multi-HMM with change points and where the number of components known

In this subsection, we assume that we have known change points as well as the number of components in the normal mixture distribution. Let w_c and m_c denote the change point set and the number of components for chromosome c. Suppose the case-control genotype data are available from the L_{cr} SNPs in the r-th region of chromosome c, $c = 1, \ldots, C$, $r = 1, \ldots, R_c$. We let $\theta_{rl}^{(c)} = 1$ indicate that SNP l from region r of chromosome c is disease-associated and $\theta_{rl}^{(c)} = 0$ otherwise. For each SNP, we first obtain a p-value by conducting a χ^2-test to assess the association between the allele frequencies and the disease status, then we convert the p-value into a z-value $Z_{rl}^{(c)}$ using the transformations proposed in Wei et al. [15] for further analysis. For chromosome c, let $\theta^{(c)} = \{\theta_{rl}^{(c)}; l = 1, \ldots, L_{cr}, r = 1, \ldots, R_c\}$ and $Z^{(c)} = \{Z_{rl}^{(c)}; l = 1, \ldots, L_{cr}, r = 1, \ldots, R_c\}$. In the following, we treat $\theta^{(c)}$ as the hidden variables and $Z^{(c)}$ as

the observed variables to consider HMMs using some assumptions.

First, we assume that the observed data are conditionally independent given the hidden states for the same region, and that different regions of the same chromosome are independent. Then we have

$$P(Z^{(c)}|\theta^{(c)}, w_c, m_c) = \prod_{r=1}^{R_c}\prod_{l=1}^{L_{cr}} P(Z_{rl}^{(c)}|\theta_{rl}^{(c)}, w_c, m_c).$$

Furthermore, let $Z_{rl}^{(c)}|\theta_{rl}^{(c)}, w_c, m_c \sim (1-\theta_{rl}^{(c)})F_{r0}^{(c)} + \theta_{rl}^{(c)}F_{r1}^{(c)}$, where $\mathcal{F}_r^{(c)} = \{F_{r0}^{(c)}, F_{r1}^{(c)}\}$ denotes the observation distribution for each SNP in region r of the chromosome c. For a non-associated SNP, we assume that the z-value distribution is a standard normal $F_{r0}^{(c)} = N(0,1)$ for all regions of chromosome c, and for a disease-associated SNP, the z-value distribution is a normal mixture, whereby $F_{r1}^{(c)} = \sum_{i=1}^{m_c} \xi_i N(\mu_{ri}^{(c)}, {\sigma_{ri}^{(c)}}^2)$, $\sum_i \xi_i = 1$, and we assume that the number of components in the normal mixture is identical for all regions of chromosome c. The normal mixture model can approximate a large collection of distributions and has been widely used elsewhere [19-21].

Second, we assume that the hidden states $\theta_r^{(c)}$ and $\theta_t^{(c)}$ are independent for the different regions, r and t. For the r-th region of chromosome c, we assume that $\theta_r^{(c)} = \{\theta_{r1}^{(c)}, \ldots, \theta_{rL_{cr}}^{(c)}\}$ is distributed as a stationary Markov chain with a transition probability $a_{rij}^{(c)} = P(\theta_{r(l+1)}^{(c)} = j|\theta_{rl}^{(c)} = i)$. Let us denote $\Lambda_{cr} = \left(a_{rij}^{(c)}\right)$ the transition matrix and $\pi_{cr} = (\pi_{r0}^{(c)}, \pi_{r1}^{(c)})'$ the stationary distribution, where $\pi_{r0}^{(c)} = a_{r10}^{(c)}/(a_{r10}^{(c)} + a_{r01}^{(c)})$ and $\pi_{r1}^{(c)} = 1 - \pi_{r0}^{(c)}$.

Let $\phi_{cr} = \{(\mu_{ri}^{(c)}, \sigma_{ri}^{(c)}, \xi_i); i = 1, \ldots, m_c\}$, then we denote $\Psi_{cr} = (\Lambda_{cr}, \pi_{cr}, \phi_{cr})$ the collection of HMM parameters for r-th region of chromosome c. When w_c and m_c are known, the maximum likelihood estimate of the HMM parameters can be obtained using the expectation-maximization (EM) algorithm [14,22].

Adaptive criterion-based partitioning (ACP) method for finding change points and the number of components

However, in practice, change points and the number of components in the normal mixture distribution are often unknown. In this subsection, for each chromosome, we will give an ACP method to conduct a model selection procedure for simultaneously finding w_c and m_c.

Candidate change point set

To effectively reduce the huge space of competing change points and save computation time, our ACP method needs a candidate change point set in advance. Here, we use a haplotype block partition method [23] to obtain the haplotype-block boundary points for each chromosome, which can be collected as the candidate change point set. Because the minimum length value of block L_{min} should be pre-specified in their haplotype block partition method, here, we let L_{min} be 300 for all our analysis.

Adaptive criterion-based partitioning procedure

Simultaneously inferring m_c and w_c can be regarded as a model selection problem. To select a desired model, the commonly used methods are established base on the criterion of minimizing the penalized negative maximum likelihood (e.g. BIC). However, many other existing criterions including BIC, assume that the observations are independent, which is not true in HMM. As a result, the effective sample sizes may be small owing to strong dependence among the observations, and the existing criterions may suffer from a failure of consistency. A data-driven penalized criterion was proposed in the Gaussian and least-squares regression model selection for independent observations [17,24,25]. Especially, [25] used this adaptive criterion for variable selection and clustering in Gaussian mixtures model and showed that this adaptive criterion outperforms other criterions (e.g. BIC) for small sample sizes. Following their work, we propose a data-driven penalized criterion for dependent observations in HMM.

Let $w_c \subset w_c^0$ denote a change point set for chromosome c, $|w_c|$ be the number of the change points in w_c, and $\Psi_c = \{\Psi_{cr}; r = 1, \ldots, R_c\}$, where w_c^0 is the candidate change point set for chromosome c. Then, we consider a penalized maximum likelihood criterion with the following form

$$\begin{aligned}\mathrm{crit}_{\lambda_c}(m_c, w_c) &= -\ln\left(P(Z^{(c)}|\hat{\Psi}_c, m_c, w_c)\right) + \mathrm{pen}_{L_c}(w_c, m_c)\\ &= -\ln\left(\sum_{\theta^{(c)}} P(Z^{(c)}, \theta^{(c)}|\hat{\Psi}_c, m_c, w_c)P(\theta^{(c)}|\hat{\Psi}_c, m_c, w_c)\right)\\ &\quad + \lambda_c D_{(m_c,w_c)},\end{aligned} \tag{1}$$

where $\hat{\Psi}_c$ is the maximum likelihood estimator of the parameters Ψ_c in HMMs for chromosome c using an EM algorithm, and the penalty function $\mathrm{pen}_{L_c}(w_c, m_c) = \lambda_c D_{(m_c,w_c)}$ is designed to avoid overfit problems. In this penalty function, $\lambda_c > 0$ is a tuning parameter to be chosen depending on sample size $L_c = \sum_{r=1}^{R_c} L_{cr}$ and $D_{(m_c,w_c)}$ is the number of parameters in the model. Furthermore, in this paper, we have $D_{(m_c,w_c)} = (|w_c|+1)D_{m_c}$, where D_{m_c} is the number of parameters that only depend on m_c. If we let $\lambda_c = \frac{\ln(L_c)}{2}$, this penalty function becomes the penalty function of BIC for HMM.

Given a value of λ_c in the penalized criterion $\mathrm{pen}_{L_c}(w_c, m_c) = \lambda_c D_{(m_c,w_c)}$, we can find $\hat{w}_c$ and $\hat{m}_c$ to minimize $\mathrm{crit}_{\lambda_c}(m_c, w_c)$ of equation 1 by running Algorithm 1.

Algorithm 1

Step 1. Input K_{max} and m_{max}

Step 2. Given the number of components $m_c = i$,

run the dynamic program (DP) algorithm to

realize the following optimal problem,

$$\hat{w}_{c,m_c=i} \longleftarrow \underset{w_c;|w_c|\le K_{max}, w_c\subset w_c^0}{\arg\min} \operatorname{crit}_{\lambda_c}(m_c = i, w_c),$$

then we can obtain a temporary change point set

$\hat{w}_{c,m_c=i}$, where $i = 1, \ldots, m_{max}$.

Step 3. Among all the temporary change point sets

$\{\hat{w}_{c,m_c=1}, \ldots, \hat{w}_{c,m_c=m_{max}}\}$ obtained in Step 2,

choose the change point set $\hat{w}_{c,m_c=\hat{m}_c}$ with the

number of components $\hat{m}_c$,

which minimizes $\operatorname{crit}_{\lambda_c}(m_c, \hat{w}_{c,m_c})$. Then let

$\hat{w}_c \longleftarrow \hat{w}_{c,m_c=\hat{m}_c}$.

In step 1, we need to give pre-specified values of K_{max} and m_{max} in advance, where K_{max} denotes the maximum value of the number of change points for each chromosome, and m_{max} is the maximum value of the number of component. As we know, the number of true change points is usually far less than the number of the candidate change points in practical applications, so we can give a smaller value for K_{max} to save computation time. For m_{max}, Wei et al. [15] suggested that values of between four and six are usually chosen.

In step 2, following the methods of [26] and [27], we provide an optimal partitioning search method for change points to estimate $\hat{w}_{c,i}$ given λ_c and m_c, which is, in essence, a dynamic program (DP) algorithm. The detailed procedure about the optimal partitioning search method is shown in Additional file 1.

However, in practice, λ_c is unknown and needs to be calibrated and estimated from the data themselves. Slope heuristics [24] as well as its generalization, dimension jump method [17], are practical and effective calibration algorithms to estimate the optimal penalty $\operatorname{pen}_{L_c,opt}(w_c, m_c) = \lambda_{c,opt} D_{(m_c,w_c)}$. Here, we propose a sliding window-based dimension jump method to estimate $\lambda_{c,opt}$, where the sliding window is used to avoid losing cases involving several successive jumps. When the width of the sliding window is 1, our proposed method becomes the dimension jump method of Wei et al. [17]. The following algorithm describes the detailed procedure for estimating $\lambda_{c,opt}$.

Algorithm 2

Step 1. Given a grid for λ_c: $0 < \lambda_{c,1} < \lambda_{c,2} < \ldots < \lambda_{c,t}$

$< \ldots < \lambda_{c,T} = \ln(L_c)$,

where $\lambda_{c,t} = \frac{t(\ln(L_c))}{T}$, $t = 1, \ldots, T$, and we set

$T = 50$ in this paper.

Step 2. For each $\lambda_{c,t}$, run Algorithm 1 to minimize

$\operatorname{crit}_{\lambda_{c,t}}(m_c, w_c)$ of equation (1),

and we can obtain $(\hat{m}_{c,t}, \hat{w}_{c,t})$.

Step 3. Given the width of sliding window $h \ge 1$,

firstly, we let set

$$\Omega_1 = \{t^*; t^* = \underset{t\in\{h+1,\ldots,T\}}{\arg\max} \{D_{(\hat{m}_{c,t-h},\hat{w}_{c,t-h})} - D_{(\hat{m}_{c,t},\hat{w}_{c,t})}\}\}$$

and $t_{end} \longleftarrow \min_{t'\in\Omega_1} t'$,

then let set

$$\Omega_2 = \{s^*; s^* \in \{t_{end} - h, \ldots, t_{end} - 1\},$$

$$D_{(\hat{m}_{c,s^*},\hat{w}_{c,s^*})} = D_{(\hat{m}_{c,t_{end}-h},\hat{w}_{c,t_{end}-h})}\}$$

and $t_{init} \longleftarrow \max_{s'\in\Omega_2} s'$.

Step 4. $\hat{\lambda}_{c,min} \longleftarrow \frac{\lambda_{c,t_{end}} + \lambda_{c,t_{init}}}{2}$

Step 5. $\hat{\lambda}_{c,opt} \longleftarrow 2\hat{\lambda}_{c,min}$

At the end of Algorithm 2, we can obtain the estimation $\hat{\lambda}_{c,opt}$ of the $\lambda_{c,opt}$. Having $\hat{\lambda}_{c,opt}$, we can then run Algorithm 1 to obtain $\hat{m}_{c,opt}$ and $\hat{w}_{c,opt}$ as well as the desired optimal model by minimizing $\operatorname{crit}_{\lambda_c}(m_c, w_c)$ of Equation (1). At the same time, we can get $\hat{\Psi}_c$, the estimates of model parameters Ψ_c based on the optimal model, where $c = 1, \ldots, C$.

Pooled FDR control procedure for different chromosomes with multiple regions

After each chromosome is divided into different regions by change points, it is desirable that that the global region-wide FDR can also be controlled by combining the test results from multiple regions of different chromosomes. In the following, we extend the chromosome-specific PLIS to the RSPLIS and operate the new procedure in three steps:

Step 1. For chromosome c ($c = 1, \ldots, C$), we search the change points to divide the whole chromosome into multiple regions using the ACP method. For each region r, we can get $\hat{\Psi}_{cr}$ by using the EM algorithm from which we can calculate the

plug-in LIS statistic $\mathrm{LIS}_{rl}^{(c)} = P_{\hat{\Psi}_{cr}}(\theta_{rl}^{(c)} = 0|Z_r^{(c)})$ for all regions of each chromosome by using the forward-backward algorithm [28].

Step 2. Combine and rank the plug-in LIS statistics from different regions of multiple chromosomes. Denote by $\mathrm{LIS}_{(1)}, \ldots, \mathrm{LIS}_{(L)}$ the ordered values and $H_{(1)}, \ldots, H_{(L)}$ the corresponding hypotheses, where $L = \sum_{c=1}^{C} \sum_{r=1}^{R} L_{cr}$.

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

We define FNR as the expected proportion of falsely accepted hypotheses. Under a compound decision-theoretic framework, the following theorem can verify that our RSPLIS is valid and asymptotically optimal. We provide the detailed proof of the theorem in Additional file 1.

Theorem 1. *Consider the multi-region HMMs defined in section 'Region-specificmulti-HMMwith change points and where the number of components known'. Let $LIS_{(1)}, \ldots, LIS_{(L)}$ be the ranked LIS values from all the regions of all chromosomes. Then, the RSPLIS procedure controls the global FDR at level α. In addition, the global FNR level of RSPLIS is $\beta^* + o(1)$, where β^* is the smallest FNR level among all valid FDR procedures at level α.*

Results

Simulation study

In this section, we design the detailed simulation studies to illustrate the performance of our ACP method in model selection; thereafter we conducted simulation studies to compare the performance of the proposed RSPLIS with that of PLIS in GWAS. All the simulations that follow were replicated 100 times.

Simulations of the ACP method performance for model selection

Simulations in this subsection were conducted to compare the performance of our ACP method with that of BIC-based partitioning (BICP) method for selecting change points and the number of components. For simplicity, we consider a single chromosome that has five stationary regions. We assume each region has the same length L_0 and set L_0 equal to 600, 900 and 1200. The detailed simulation parameter settings are given in Table 1. With different parameter settings, we expect that the first two change points can be identified easily, while the last two change points are harder to be identified.

In this simulation, we give the candidate change point set

$$w^0 = \{300i; i = 1, 2, \ldots, \frac{5L_0 - 300}{300}\},$$

which ensures that the true change point set $\{iL_0; i = 1, \ldots, 4\} \subset w^0$. To compare the performance of the two methods, we used sensitivity and specificity as measures. Sensitivity is defined as the average proportions of the true change points which are correctly identified as change points over the 100 times and the specificity is defined as the average proportions of the false change points which are not identified as the change points over the 100 times. We set $K_{max} = 8$ and $m_{max} = 5$ for our ACP method. At the same time, h takes the values of 2, and 20 for the window.

The simulation results are summarized in Tables 2 and 3. From these tables, we can see that BICP misses most of the true change points and the true number of components. Moreover, we can also see our ACP has

Table 1 Parameter settings of simulated data

Region	Transition matrix	Stationary distribution	z-value distribution (Null)	z-value distribution (No-null)
1	$\begin{pmatrix} 0.98 & 0.02 \\ 0.20 & 0.80 \end{pmatrix}$	(0.91, 0.09)	$N(0, 1)$	$0.1N(1.0, 1) + 0.9N(3.0, 1)$
2	$\begin{pmatrix} 0.98 & 0.02 \\ 0.95 & 0.05 \end{pmatrix}$	(0.98, 0.02)	$N(0, 1)$	$0.8N(1.5, 1) + 0.2N(4.5, 1)$
3	$\begin{pmatrix} 0.98 & 0.02 \\ 0.45 & 0.55 \end{pmatrix}$	(0.96, 0.04)	$N(0, 1)$	$0.4N(1.5, 1) + 0.6N(3.5, 1)$
4	$\begin{pmatrix} 0.98 & 0.02 \\ 0.15 & 0.85 \end{pmatrix}$	(0.88, 0.12)	$N(0, 1)$	$0.2N(1.0, 1) + 0.8N(3.0, 1)$
5	$\begin{pmatrix} 0.98 & 0.02 \\ 0.15 & 0.85 \end{pmatrix}$	(0.88, 0.12)	$N(0, 1)$	$0.2N(1.0, 1) + 0.8N(3.0, 1)$

Table 2 The results of comparing ACP method with BICP method for selecting the true number of components $m = 2$

Length of region	Measures	ACP			BICP
(L)		$h = 2$	$h = 10$	$h = 20$	
600	Sensitivity	0.86	0.83	0.21	0.18
	Specifity	0.97	0.96	0.80	0.80
900	Sensitivity	0.83	0.80	0.13	0.16
	Specifity	0.96	0.96	0.78	0.79
1200	Sensitivity	0.81	0.78	0.12	0.15
	Specifity	0.95	0.94	0.78	0.79

higher specificity than BICP. The reason for the poorer behavior of BICP may be related to the lack of independent observations in this experiment, so there may be a smaller effective sample size for BIC. In addition, based on the simulation results, we can see that the performance of our ACP is very good for $h = 2$ and $h = 10$, but is very poor for $h = 20$, so we suggest h should not be more than 10 in practice.

HMM-based simulations for comparing RSPLIS with PLIS in GWAS

For simplicity, in this simulation, suppose there are two chromosomes ($c = 2$) in total, each of which consists of two stationary regions, and each region has 2000 SNPs ($L_{cr} = 2000, r = 1, 2, c = 1, 2$). For each chromosome, we set $K_{max} = 4$, $m_{max} = 3$, and $h = 2$ for our ACP method and gave the candidate change point set $w_c^0 = \{300i; i = 1, 2, \ldots, \frac{L_c - 300}{300}\}$, $c = 1, 2$. The purpose of this simulation is to compare RSPLIS with PLIS by finding disease-associated SNPs while controlling the FDR at a pre-specified level $\alpha = 0.1$ for the two chromosomes (combining chromosomes 1 and 2). We conducted simulation studies in the following two cases.

Case 1. In this case, we varied the dependence parameters in transition matrices of HMM and kept the other parameters fixed, and then we investigated the behavior of RSPLIS and compare it with PLIS procedure to identify casual SNPs at the different disease risk levels. We used the parameter settings in Table 4, where we varied the degree of dependence among SNPs in region 2 by changing the value of υ_1 ($\upsilon_1 = 0, 0.15, 0.30, 0.45$). Furthermore, we let $\mu_{21}^{(c)} = \mu_{11}^{(c)} + 1.5$, $\mu_{11}^{(2)} = \mu_{11}^{(1)} + 0.5$ and varied the disease risk parameter $\mu_{11}^{(1)}$ from 0.5 to 2.0 with an increment of 0.5.

Table 3 The results of comparing ACP method with BICP method for selecting the true change point set $\{iL; i = 1, \ldots, 4\}$

Length of region	Measures	ACP			BICP
(L)		$h = 2$	$h = 10$	$h = 20$	
600	Sensitivity	0.91	0.90	0.27	0.22
	Specifity	0.90	0.85	0.74	0.73
900	Sensitivity	0.68	0.67	0.13	0.13
	Specifity	0.86	0.85	0.81	0.83
1200	Sensitivity	0.42	0.29	0.08	0.08
	Specifity	0.89	0.82	0.86	0.87

Case 2. In contrast to Case 1, to assess the performance of RSPLIS at the different disease risk levels, we varied the parameters of the z-value distribution while fixing the other parameters. We used the parameter settings in Table 5, where we varied the parameters of the z-value distribution by changing the value of υ_2 ($\upsilon_2 = 0.5, 1, 1.5, 2$). Furthermore, we let $\mu_{21}^{(c)} = \mu_{11}^{(c)} + \upsilon_2$, $\mu_{11}^{(2)} = \mu_{11}^{(1)} + 0.5$ and varied the disease risk parameter $\mu_{11}^{(1)}$ from 0.5 to 2.0 with an increment of 0.5.

The simulation results are shown in Figures 1, 2, 3 and 4. From Figure 1 and Figure 3, we can obviously see that both RSPLIS and PLIS are well controlled at FDR level 0.1 asymptotically. Figure 2 and Figure 4 inform us that PLIS is dominated by RSPLIS for the power at Case 1 and Case 2, which indicates that our RSPLIS procedure is effective at dividing the chromosomes into smaller and more homogeneous regions by searching the change points. In addition, the difference in FNR levels (RSPLIS vs. PLIS) becomes smaller as $\mu_{11}^{(1)}$ increases for each model, which implies that RSPLIS is especially useful when the disease signals are weak.

To show that the higher power of RSPLIS is not gained to the detriment of a higher FDR level, we conducted a further simulation study. This study evaluated the sensitivities at different FDR levels for $\upsilon_1 = 0, 0.15, 0.30, 0.45$ in Case 1, and $\upsilon_2 = 0.5, 1.0, 1.5, 2.0$ in Case 2, where the sensitivities were calculated as the average proportions of correctly identified SNPs over the 100 replications. For the purpose of illustration, we have only listed the results for $\upsilon_1 = 0.15$ of Case 1 and $\upsilon_2 = 1.0$ of Case 2 in Figure 5 and Figure 6 respectively, because the other results were broadly similar. It is clear from Figures 5 and 6 that RSPLIS discovers more true disease-associated SNPs than PLIS at the same FDR level.

Genotype-based simulations for comparing RSPLIS with PLIS

This simulation evaluated the performance of selecting the *relevant* SNPs for RSPLIS and PLIS based on the genotype data set. In contrast to the simulation study in Subsection 'Application to the Daly data set', we generated case-control genotype data with more realistic LD patterns. To this end, we constructed a genotype pool

Table 4 Parameter settings of simulated data for Case 1

Chromosome (c)	Region (r)	Transition matrix	z-value distribution (Null)	z-value distribution (Non-null)
1	1	$\begin{pmatrix} 0.98 & 0.02 \\ 0.03 & 0.97 \end{pmatrix}$	$N(0,1)$	$N(\mu_{11}^{(1)},1)$
	2	$\begin{pmatrix} 0.98 & 0.02 \\ 0.03+\upsilon_1 & 0.97-\upsilon_1 \end{pmatrix}$	$N(0,1)$	$N(\mu_{21}^{(1)}=\mu_{11}^{(1)}+1.5,1)$
2	1	$\begin{pmatrix} 0.98 & 0.02 \\ 0.05 & 0.95 \end{pmatrix}$	$N(0,1)$	$N(\mu_{11}^{(2)}=\mu_{11}^{(1)}+0.5,1)$
	2	$\begin{pmatrix} 0.98 & 0.02 \\ 0.05+\upsilon_1 & 0.95-\upsilon_1 \end{pmatrix}$	$N(0,1)$	$N(\mu_{21}^{(2)}=\mu_{11}^{(2)}+2.0,1)$

composed of genotypes from 60 samples for 23 chromosomes by randomly matching the pair of the known phased 120 haplotypes from the Illumina 550K. For simplicity, we only used SNPs selected from 2001-th SNP to 6000-th SNP of chromosome 7 and chromosome 15, respectively. Four SNPs were selected from each chromosome as the disease causal SNPs, each with a relative risk of 1.5. Specifically, the four SNPs, 400-th, 900-th, 1750-th, 3200-th, were chosen to be far away on chromosome 7, while the four other SNPs, 5600-th, 5604-th, 5608-th, 5612-th, were chosen based on their proximity to each other (i.e., separated by 3 SNPs) on chromosome 15.

For each subject, we first obtained the genotype, X, by drawing a genotype from the genotype pool at random. Using genotype X, we then simulated the disease status, Y, of this subject using the logistic regression model,

$$P(Y=1|X)=\frac{\exp(\beta_0+\sum_{i=1}^{8}\beta_i X_i)}{1+\exp(\beta_0+\sum_{i=1}^{8}\beta_i X_i)},$$

where $\beta_0 = -9.5425$ for a disease rate of 0.03, $\beta_i = \log(1.5)$, for $i = 1,\ldots,8$. We repeated the sampling procedure until we obtained 1000 cases and 1000 controls are obtained. The eight disease casual SNPs were then removed from the simulated data set, and the 39 SNPs that contained the three adjacent SNPs on each side of the eight disease-causal SNPs were regarded as the relevant SNPs. We assessed the performance of the testing procedure by the selection rate of *relevant* SNPs, where the percentages of the true positives (sensitivity) selected by the top M SNPs could be calculated easily. We set $m_{max} = 6$, $h = 2$, and $K_{max} = 5$ for the ACP method.

We have plotted the average sensitivity curves for comparisons of RSPLIS vs. PLIS in Figure 7. It is apparent that our RSPLIS dominates PLIS in ranking the *revelant* SNPs. In summary, these results show that exploiting the heterogeneous chromosomal regions and searching the change points to find chromosomal regions that are more homogeneous has improved the precision of RSPLIS in that the number of false positives has been reduced while the statistical power has increased.

Table 5 Parameter settings of simulated data for Case 2

Chromosome (c)	Region (r)	Transition matrix	z-value distribution (Null)	z-value distribution (Non-null)
1	1	$\begin{pmatrix} 0.98 & 0.02 \\ 0.025 & 0.975 \end{pmatrix}$	$N(0,1)$	$N(\mu_{11}^{(1)},1)$
	2	$\begin{pmatrix} 0.98 & 0.02 \\ 0.025 & 0.975 \end{pmatrix}$	$N(0,1)$	$N(\mu_{21}^{(1)}=\mu_{11}^{(1)}+\upsilon_2,1)$
2	1	$\begin{pmatrix} 0.98 & 0.02 \\ 0.025 & 0.975 \end{pmatrix}$	$N(0,1)$	$N(\mu_{11}^{(2)}=\mu_{11}^{(1)}+0.5,1)$
	2	$\begin{pmatrix} 0.98 & 0.02 \\ 0.025 & 0.975 \end{pmatrix}$	$N(0,1)$	$N(\mu_{21}^{(2)}=\mu_{11}^{(2)}+\upsilon_2,1)$

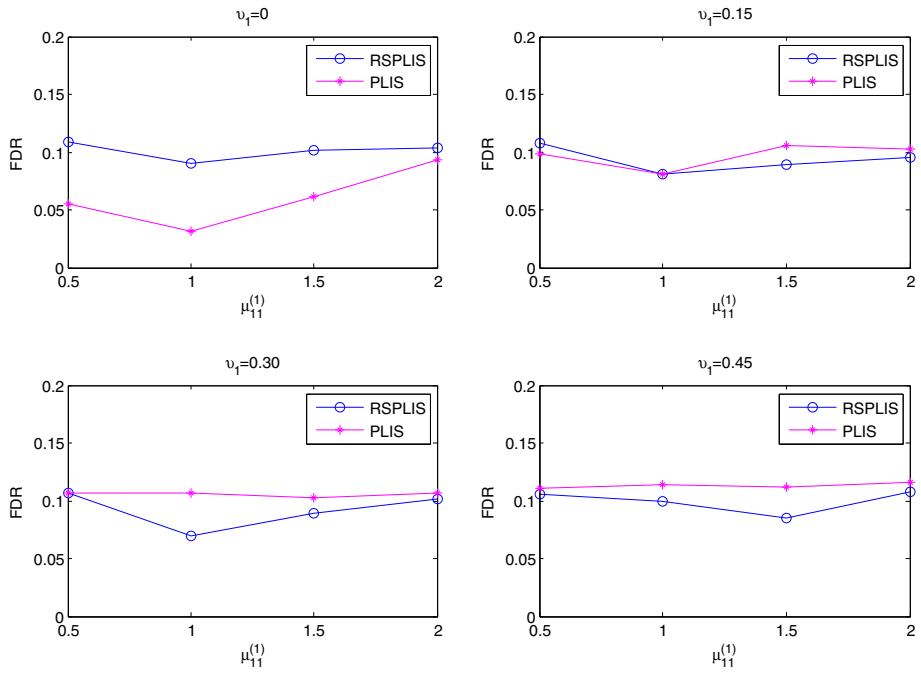

Figure 1 Shows that the FDR levels of the two methods are controlled at the level 0.10 asymptotically for four different parameter settings in Case 1.

Application to the Daly data set

The data are derived from the 616 kilobase region of human chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease as determined by genotyping 103 common SNPs (>0.05 minor allele frequency) for 129 trios [18]. All offspring belong to the case population, while almost all parents belong to the control population. In the entire data, there are 144 cases and 243 controls. Daly et al. [18] have also shown that there are 11 blocks and strong LD between the SNPs and their neighboring SNPs in each block.

Model selection and estimation of HMM parameters

First, we used the *K*-Nearest neighbor method proposed by the R package [29] to impute the missing genotypes

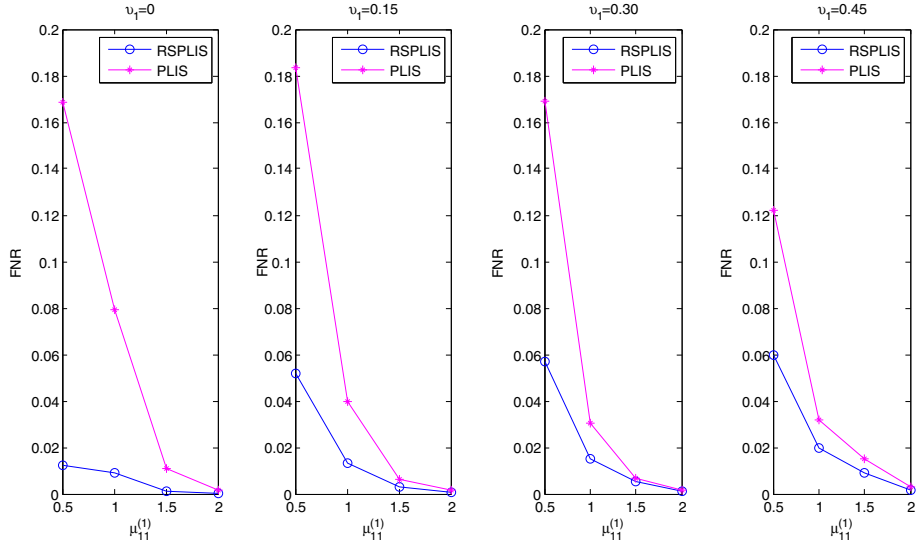

Figure 2 Shows the FNR of PLIS is much higher than that of RSPLIS for four different parameter settings in Case 1.

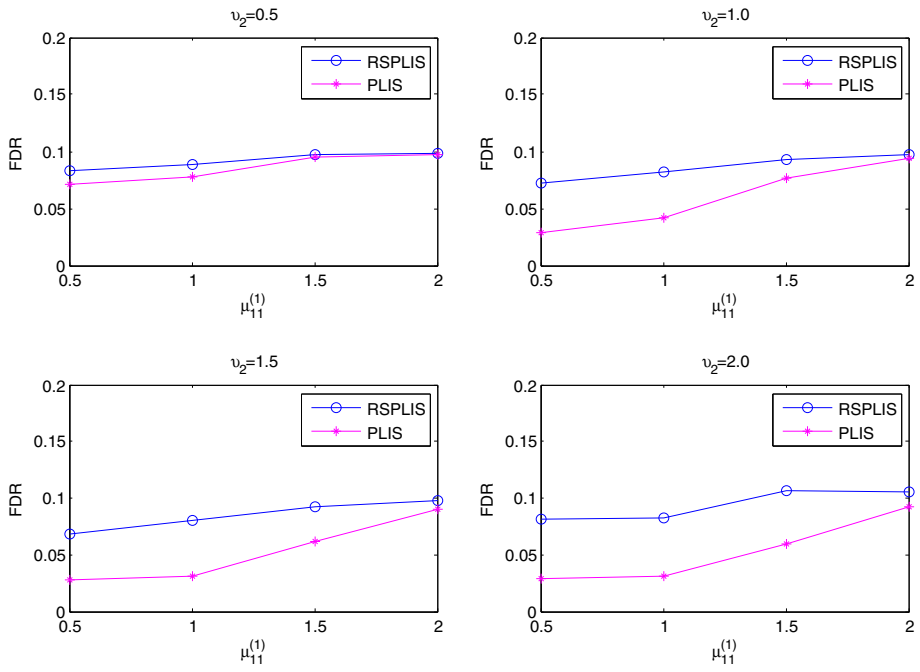

Figure 3 Shows the FDR levels of the two methods are controlled at 0.10 asymptotically for four different parameter settings in Case 2.

from the Daly et al. data [18]. Then, for each SNP, we obtained a p-value by conducting a χ^2-test to assess associations between the allele frequencies and the disease status, furthermore, we get z-value by transforming p-value. We assumed that the null distribution is standard normal $N(0,1)$ and the non-null distribution is a normal mixture $\sum_{i=1}^{m_c} \xi_i N(\mu_{ri}^{(c)}, {\sigma_{ri}^{(c)}}^2)$, where $c = 1$ because there is only one chromosome in the Daly et al. data [18]. We used our ACP method to select the number of components and change points, where the parameters in HMMs were estimated by the EM algorithm. Thereafter, RSPLIS was used for multiple testing.

Data analysis

Because the Daly et al. data [18] have only 103 SNPs, we assumed only one change point for these data in our analysis. Thus, we took $L_{min} = 20$, $m_{max} = 6$, $h = 2$ and

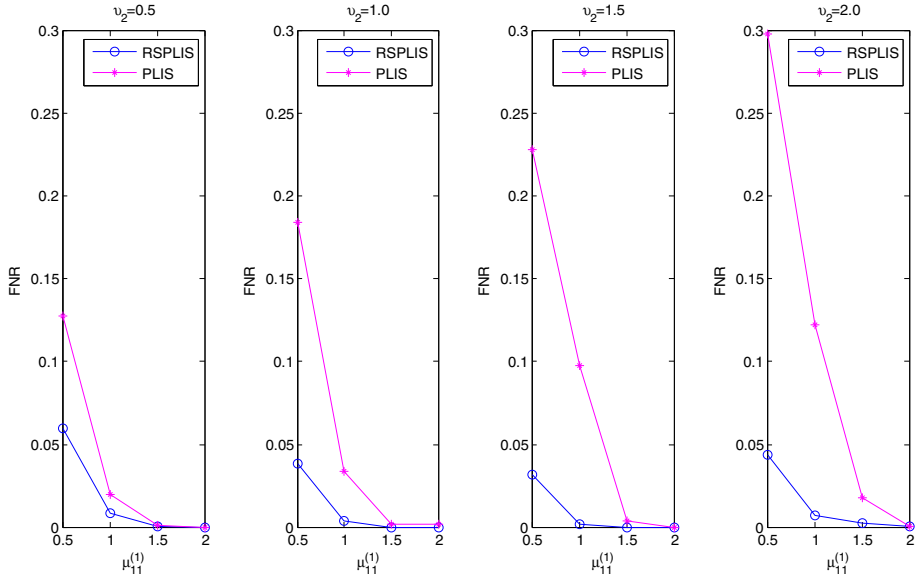

Figure 4 Shows the FNR of PLIS is much higher than that of RSPLIS for four different parameter settings in Case 2.

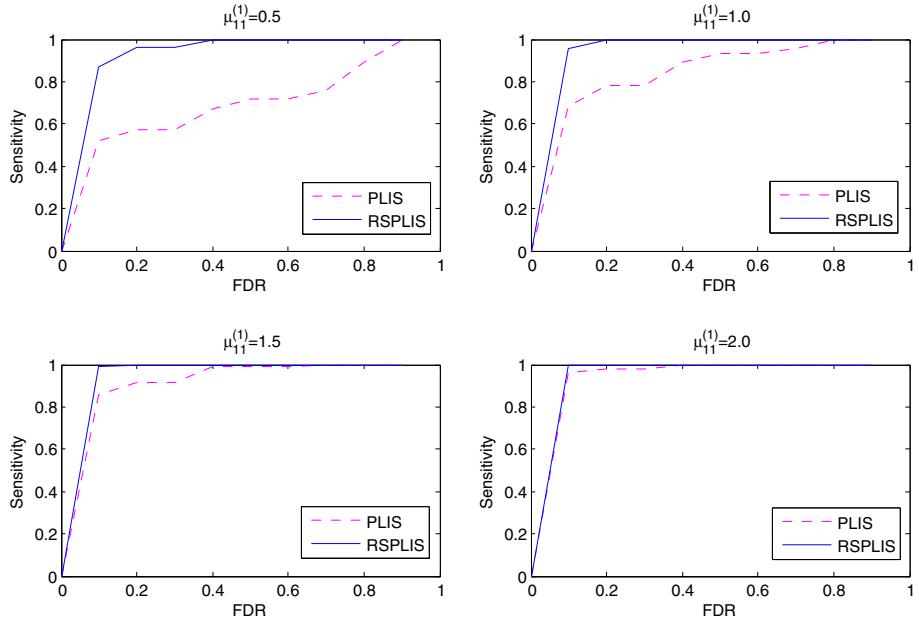

Figure 5 Shows the ranking efficiency for $v_1 = 0$ in Case 1: RSPLIS has higher sensitivity than PLIS at the same FDR level; there is a more dramatic improvement when the signals are weak ($\mu_{11}^{(1)}$ is small).

took $K_{max} = 1$ for our ACP method. For the purpose of comparison, we also used the PLIS method to analyze the Daly et al. data [18]. The data [18] were collected to identify genetic variants conferring susceptibility to Crohn's disease and nine SNPs were identified [30]. For the purpose of illustration, we only list here the LIS statistics and LIS ranks for the nine casual SNPs (Table 6). Based on the definition of LIS statistic given in Section 'Pooled FDR control procedure for different chromosomes with multiple regions', it is obvious that the smaller value of the LIS statistic means a larger probability that this SNP is associated with the disease. From Table 6, the rankings for eight causal SNPs illustrate that RSPLIS offers a marked improvement over PLIS, with the exception of locus 73. It

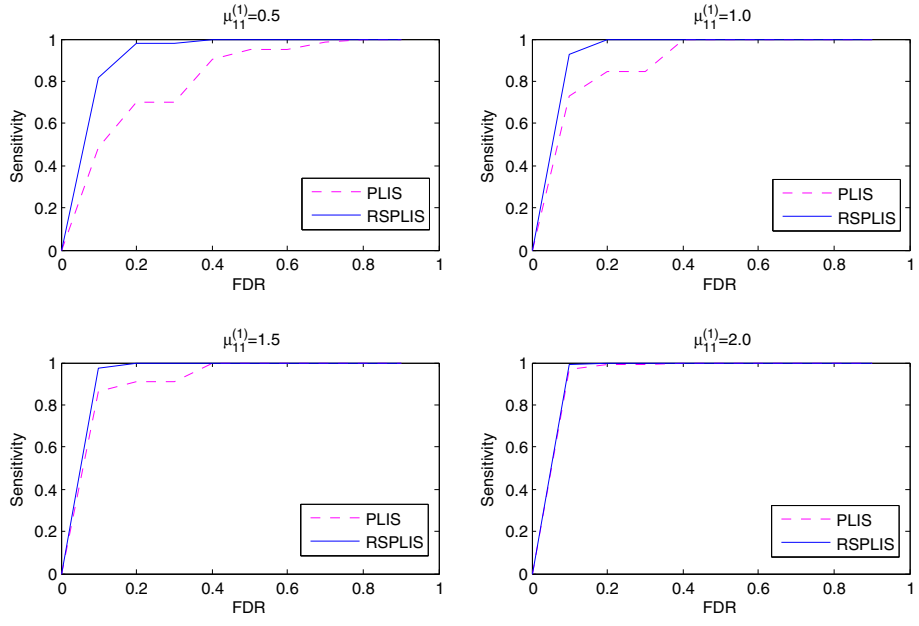

Figure 6 Shows the ranking efficiency for $v_1 = 0$ in Case 2: RSPLIS has higher sensitivity than PLIS at the same FDR level; there is a more dramatic improvement when the signals are weak ($\mu_{11}^{(1)}$ is small).

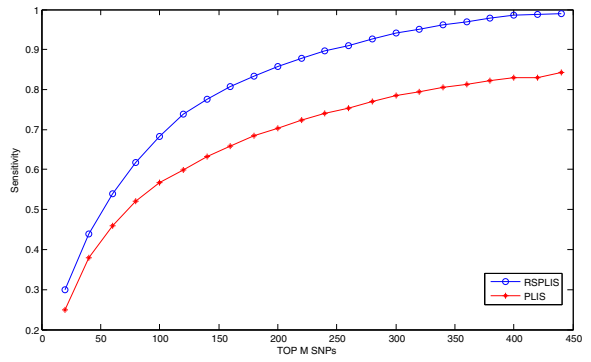

Figure 7 The four SNPs in Chr 1 are far away from each other, while the four in Chr 2 are separated by only three SNPs. We calculated the sensitivity selected by the top M SNPs, ranked by RSPLIS and by PLIS.

is not surprising that not all of the nine causal SNPs are top ranked because non-casual SNPs that are strongly linked to the casual SNP may also be top-ranked. In summary, we can see that our method not only makes better rankings but also has smaller values for LIS statistics for most of the true casual SNPs. In addition, Table 6 shows that the LIS values obtained using RSPLIS are far lower than those obtained using PLIS. The reason may be that each region found by the ACP method has a smaller sample size for statistical inference in HMM, so this may affect the values obtained for the LIS statistic.

Discussion

Large-scale multiple testing under dependence is holding promise in identifying genetic variants for GWAS. Previous research has focused on large-scale multiple testing under a HMM for a single chromosome ([14,15]). In the present paper, we extended chromosome-specific PLIS to RSPLIS to analyze SNP data arising from large-scale GWAS by an adaptive penalized criterion. By dividing the whole chromosome into more homogeneous regions and conducting the extended pooling dependent testing procedure, we showed that the accuracy of a multiple testing procedure was improved when there are multiple change points along the whole chromosome. The numerical performances of our RSPLIS procedure were investigated using both simulated studies and real data analysis. The results showed that RSPLIS is more powerful than PLIS at identifying small effects in GWAS.

Table 6 Results of PLIS and RSPLIS for the known 9 casual SNPs in Daly data, which were reported as significant (Rioux et al., 2001)

SNP name	SNP location	LIS statistics		LIS ranks	
	(th)	RSPLIS	PLIS	RSPLIS	PLIS
IGR2055a-1	25	1.5994e-18	7.5411e-04	5	7
IGR2060a-1	26	1.7388e-18	8.5125e-04	7	9
IGR2063b-1	27	1.7389e-18	8.4605e-04	6	8
IGR2096a-1	33	8.0144e-18	4.1287e-03	39	51
IGR2198a-1	38	1.0731e-17	5.3388e-03	56	79
IGR2230a-1	48	3.4931e-18	2.1243e-03	11	13
IGR3081a-1	73	7.6344e-18	3.9276e-03	37	34
IGR3096a-1	77	3.7265e-18	2.3153e-03	12	14
IGR3236a-1	92	1.3554e-18	5.9574e-04	3	5

However, our method could be improved in several ways. In the present paper, we conducted large-scale multiple testing under a special form of dependence (HMM) for the hypotheses. Because complex LD structure(s) are usually stored in SNP data, the Markov chain may not be the most appropriate model for SNP dependence. Therefore, general forms of dependence such as the Markov random field should be considered in future, where the whole network is divided into a region-specific Markov random field network; this would improve the screening efficiency in GWAS.

Besides, the question of how to select an ideal candidate change point set is one issue that needs further consideration. Clearly, better prior knowledge can help us find the change point set to reduce the space of competing models in the model selection procedure. Thus, a better algorithm needs to be developed by using prior information to obtain the candidate change point set.

The computational complexity and feasibility of our RSPLIS approach for analyzing GWAS data that contain tens of thousands of SNPs merit further discussion. The RSPLIS method is made up of three independent procedures: a procedure for getting the candidate change point set, the adaptive criterion-based model selection procedure with HMM parameter estimations, and the pooled FDR control procedure for all the chromosomes with multiple regions, where the second procedure is the most time consuming. Fortunately, our method runs the second procedure chromosome-by-chromosome, which facilitates parallel computing. For each chromosome, say, the c-th chromosome, the computational complexity for three procedures is $O(L_{min}^2 L_c)$, $O(8TK_{max}^2 m_{max}|w_c^0|L_c)$, and $O(L_c \log(L_c))$ respectively, where L_c denotes the number of SNPs on the c-th chromosome, $|w_c^0|$ denotes the number of the change points in candidate change point set w_c^0, $T = 50$ in our Algorithm 2, and m_{max} is usually chosen between four and six [15]. In addition, our method is very flexible, which allows users to set the minimum length value of block (L_{min}) as well as the maximum value of the number of change points (K_{max}) in large-scale GWAS. Based on our simulation studies, with the setting, $L_{min} = 300$, $K_{max} = 5$, $T = 50$, and $m_{max} = 6$, it took about 40 min for our RSPLIS procedure to analyze 8000

SNPs from two chromosomes. We expect that the running time for large-scale GWAS is still acceptable because we can use parallel computing for each chromosome.

Conclusions

In this paper, we first modeled the observed dependent SNP data via region-specific multiple HMMs divided by change points, where we developed a novel data-driven penalized criterion combined with the DP algorithm to find change points. Second, we proposed a RSPLIS method to conduct the dependent tests from multiple chromosomes with different regions for GWAS. Finally, we have shown the numerical performances of the RSPLIS procedure using both simulated studies and analysis of a real data set.

Availability

Matlab and R code for RSPLIS can be accessed at http://math.nenu.edu.cn/faculty/wszhu/softwares/RSPLIS.html. This site contains the program files and code introduction.

Additional file

Additional file 1: The derivation of the dynamic program algorithm and the proof of theorem 1. Additional file 1 contains the derivation of the dynamic program (DP) algorithm for the step 2 in Algorithm 1, the derivation of the RSPLIS procedure and proof of theorem 1.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Designed the experiments: JX, WZ; Performed the experiments: JX; Wrote the paper: JX, WZ and JG. All authors contributed to the analysis, read, and approved the final manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (no. 11025102, 11001044 and 11371083); the Fundamental Research Funds for the Central Universities (no. 11CXPY007, 10JCXK001); Natural Science Foundation of Jilin Province (no. 201215007, 20100401); Scientific Research Foundation of Returned Scholars, MOE of China; Program for Changjiang Scholars and Innovative Research Team in University.

Received: 5 June 2013 Accepted: 20 September 2013
Published: 25 September 2013

References

1. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57:**289–300.
2. Efron B, et al.: **Empirical bayes analysis of a microarray experiment.** *J Am Stat Assoc* 2001, **96:**1151–1160.
3. Miller C, et al.: **Controlling the false-discovery rate in astrophysical data analysis.** *Astronomical J* 2001, **122:**3492–3505.
4. Tusher V, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Nat Acad Sci* 2001, **98:**5116–5121.
5. Storey J, Tibshirani R: **Statistical significance for genome-wide studies.** *Proc Nat Acad Sci* 2003, **100:**9440–9445.
6. Dudoit S, et al.: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Stat Sinica* 2002, **12:**111–139.
7. Sabatti C, Service S, Freimer N: **False discovery rate in linkage and association genome screens for complex disorders.** *Genetics* 2003, **164:**829–833.
8. Meinshausen N, Rice J: **Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses.** *Ann Stat* 2006, **34:**373–393.
9. Schwartzman A, Dougherty R, Taylor J: **False discovery rate analysis of brain diffusion direction maps.** *Ann Stat* 2008, **2:**153–175.
10. Royle JP, Dykstra RL: **A method for finding projection onto Guo, W., and Peddada, S. (2008), Adaptive choice of the number of bootstrap samples in large scale multiple testing.** *Stat Appl Genet Mol Biol* 2008, **7**(1):13.
11. Sabatti C: **Genomewide association analysis of metabolic phenotypes in a birth cohort from a founder population.** *Nat Genet* 2009, **41:**35–46.
12. Wei Z, Li H: **A Markov random field model for network-based analysis of genomic data.** *Bioinformatics* 2007, **23:**1537–1544.
13. Wei Z, Li H: **A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression eata.** *Ann Appl Stat* 2008, **2:**408–429.
14. Sun W, Cai T: **Large-scale multiple testing under dependence.** *J R Stat Soc Ser B* 2009, **71:**393–424.
15. Wei Z, Sun W, Wang K, Hakonarson H: **Multiple testing in genome-wide association studies via hidden Markov models.** *Bioinformatics* 2009, **25**(21):2802–2808.
16. Li H, Wei Z, Maris J: **A hidden Markov random field model for genome-wide association studies.** *Biostatistics* 2010, **11**(1):139–150.
17. Arlot S, Massart P: **Data-driven calibration of penalties for least-squares regression.** *J Mach Learn Res* 2009, **10:**245–279.
18. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: **High-resolution haplotype structure in the human genome.** *Nat Genet* 2001, **29:**229–232.
19. Magder L, Zeger S: **A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians.** *J Am Stat Assoc* 1996, **91:**1141–1151.
20. Pan W, Lin J, Le CT: **A mixture model approach to detecting differentially expressed genes with microarray data.** *Funct Integr Genomics* 2003, **3:**117–24.
21. Efron B: **Large-scale simultaneous hypothesis testing: the choice of a null hypothesis.** *J Am Stat Assoc* 2004, **99:**96–104.
22. Ephraim Y, Merhav N: **Hidden Markov processes.** *IEEE Trans Inf Theory* 2002, **48:**1518–1569.
23. Zhao Y, Xu Y, Wang Z, Zhang H, Chen G: **A better block partition and ligation strategy for individual haplotyping.** *Bioinformatics* 2008, **24**(23):2720–2725.
24. Birge L, Massart P: **Minimal penalties for gaussian model selection.** *Probability Theory Relat Fields* 2007, **138**(1–2):33–73.
25. Maugis C, Michel B: **Slope heuristics for variable selection and clustering via Gaussian mixtures.** *Tech Rep* 2008. 6550,INRIA.
26. Yao Y: **Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches.** *Ann Stat* 1984, **12**(4):1434–1447.
27. Jackson B, Sargle JD, Barnes D, Arabhi S, Alt A, Gioumousis P, Gwin E, Sangtrakulcharoen P, Tan L, Tsai TT: **An algorithm for optimal partitioning of data on an interval.** *IEEE Signal Process Lett* 2005, **12**(2):105–108.
28. Rabiner L: **A tutorial on hidden markov models and selected applications in speech recognition.** *Proc IEEE* 1989, **77:**257–286.
29. Schwender H, Ickstadt K: **Imputing missing genotypes with weighted k nearest neighbors.** *J Toxicol Environ Health, Part A* 2012, **75:**438–446.
30. Rioux JD, Daly MJ, Silverberg M, Lindblad K, Steinhart H, et al.: **Genetic variation in the 5q31 cytokine gene cluster studconfers susceptibility to Crohn disease.** *Nat Genet* 2001, **29:**223–228.

doi:10.1186/1471-2105-14-282
Cite this article as: Xiao *et al.*: Large-scale multiple testing in genome-wide association studies via region-specific hidden Markov models. *BMC Bioinformatics* 2013 **14**:282.