BMC Genetics

METHODOLOGY ARTICLE Open Access

Network-based regularization for high dimensional SNP data in the case–control study of Type 2 diabetes

Jie Ren[1], Tao He[2], Ye Li[3], Sai Liu[4], Yinhao Du[1], Yu Jiang[5] and Cen Wu[1*]

Abstract

Background: Over the past decades, the prevalence of type 2 diabetes mellitus (T2D) has been steadily increasing around the world. Despite large efforts devoted to better understand the genetic basis of the disease, the identified susceptibility loci can only account for a small portion of the T2D heritability. Some of the existing approaches proposed for the high dimensional genetic data from the T2D case–control study are limited by analyzing a few number of SNPs at a time from a large pool of SNPs, by ignoring the correlations among SNPs and by adopting inefficient selection techniques.

Methods: We propose a network constrained regularization method to select important SNPs by taking the linkage disequilibrium into account. To accomodate the case control study, an iteratively reweighted least square algorithm has been developed within the coordinate descent framework where optimization of the regularized logistic loss function is performed with respect to one parameter at a time and iteratively cycle through all the parameters until convergence.

Results: In this article, a novel approach is developed to identify important SNPs more effectively through incorporating the interconnections among them in the regularized selection. A coordinate descent based iteratively reweighed least squares (IRLS) algorithm has been proposed.

Conclusions: Both the simulation study and the analysis of the Nurses's Health Study, a case–control study of type 2 diabetes data with high dimensional SNP measurements, demonstrate the advantage of the network based approach over the competing alternatives.

Keywords: Case–control association study, Network-based regularization, Regularized logistic regression, Type 2 diabetes, Variable selection

Background

Type 2 diabetes mellitus (T2D), a chronic metabolic disorder, has been a major public health concern for years. An estimated 366 million cases of T2D over the world are expected by the year 2030 [1]. To better understand T2D etiology, significant efforts have been devoted to the identification of genetic markers that may contribute to the predisposition of the disease. The large scale genome–wide association studies (GWAS) has proven to be powerful in finding the association between individual genetic variant (like SNPs) and complex diseases, including type 2 diabetes. However, those identified SNPs from existing studies can only account for about 10% of the genetic variance of type 2 diabetes [2], which motivate the development of more advanced statistical methodologies with the hope to explain the missing heritability.

One major limitation shared by many of the previous studies, especially the early ones, is that they are marginal in the sense that one or a small number of genetic factors are analyzed at a time. Since complex disease phenotypes are associated with the joint effects of multiple genetic factors, signals with weak or moderate marginal but strong joint effects may not be captured by the marginal analysis.

As unprecedented amount of high dimensional omics data has been generated from high–throughput profiling studies, extensive regularized variable selection methods

* Correspondence: wucen@ksu.edu
[1]Department of Statistics, Kansas State University, 1116 Mid-Campus Drive N., 66506 Manhattan, KS, USA
Full list of author information is available at the end of the article

such as LASSO [3] and elastic net [4], have been proposed to identify genes that are associated with disease phenotypes, with the genes being treated as variables. More recently, to incorporate the interconnection information, or network structure existing among genetic variants into the selection procesure, the network–constrained regularization approaches have been developed, as in Li and Li [5] and Huang et al. [6], among many others. In particular, Huang et al. [6] developed the sparse Laplacian shrinkage (SLS) penalty built upon the combination of MCP (Zhang [7]) and Laplacian quadratic associated with a graph. They also demonstrated that in high dimension settings with $p \gg n$ under reasonable assumptions, SLS is selection consistent and equivalent to the oracle Laplacian shrinkage estimator with high probability.

This study has been partially motivated by analyzing the case control data from the Nurses's Health Studies (NHS) and studies alike. As a major component of the Gene Environment Association Studies Initiative, NHS was launched in 1976 in order to identify important genetic variants related to type 2 diabetes and gene–trait association under environmental exposures [8]. To accommodate the linkage disequilibrium (LD) existing among SNPs, we adopt a network measure and incorporate it in SLS. We further extend the SLS into the penalized logistic regression model for the analysis of the T2D case control data, and develop an efficient coordinate descent based algorithm. Compared with the alternatives, the proposed method can borrow strength from the correlation among SNPs and leads to more meaningful identification of important ones.

We first introduce the data and model settings, and describe the proposed approach. An efficient computational algorithm is subsequently developed. Simulation study demonstrates the significant advantage of the proposed approach over multiple competing alternatives. We analyze NHS type 2 diabetes data with high dimensional SNP measurements.

Methods

Denote the i^{th} subject by using the subscript *i*. Let (X_i, Y_i), $i = 1, \ldots, n$ be n independent and identically distributed random vectors. Y_i is the binary response variable where $y_i = 1$ indicating the case of disease, and 0 otherwise. X_i is the p–dimensional design vector of SNPs. Assuming that y_i follows a binomial distribution, then

$$P(y_i = 1|\eta_i) = \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

where η_i is the i^{th} component of $\eta = X\beta$, and $\beta = (\beta_i, \ldots, \beta_p)^T$ is the regression coefficient vector. The corresponding loss function is the negative log-likelihood

$$L(\eta) = \frac{1}{n}\sum_{i=1}^{n} L_i(\eta_i) = -\frac{1}{n}\sum_{i=1}^{n} \log P(Y_i = y_i|\eta_i) \qquad (1)$$

Regularized logistic regression

As the T2D disease status is only affected by a small number of important SNPs that are associated with the disease, and the dimensionality of the total number of SNPs is much larger than the sample size n, the problem is of a "large p, small n" nature. Regularization is a natural tool for such type of problem appropriate in both biological and statistical sense. By imposing penalty function to the loss function in (1), we have the following penalized likelihood

$$Q(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i \log\pi_i + (1-y_i)\log(1-\pi_i)\} + P(\beta;\lambda,\gamma) \qquad (2)$$

where $P(\beta;\lambda,\gamma)$ is the penalty function with tuning parameters λ and γ. A seemingly straightforward choice for the penalty is

$$P(\beta;\lambda,\gamma) = \sum_{m=1}^{p} \rho(\beta_m;\lambda_1,\gamma)$$

where $\rho(t;\lambda_1,\gamma) = \lambda_1 \int_0^{|t|} \left(1 - \frac{x}{\gamma\lambda_1}\right)_+ dx$ is the MCP penalty with tuning parameter λ_1 and regularization parameter γ (Zhang [7]).

For the SNPs, MCP is imposed on their regression coefficients. Penalized regression will shrink some components of coefficient vector β to zero, which indicates that the corresponding SNPs are not associated with the disease status y. SNPs with nonzero coefficients are treated as important variants. A major limitation of MCP here is that it ignores the interconnections among SNPs, while the high correlation among genetic variants, including SNPs, have been widely observed and reported due to LD. We use a network structure to describe the correlation pattern among SNPs. In a SNP network, a node corresponds to a SNP, and if the two SNPs are statistically or biologically associated, the two corresponding nodes are connected. To incorporate the network information, we adopt the sparse Laplacian penalty from Huang et al. [6] as follows:

$$P(\beta;\lambda,\gamma) = \sum_{m=1}^{p} \rho(\beta_m;\lambda_1,\gamma) + \lambda_2 \sum_{1 \le m < k \le p} |a_{mk}| \left[\beta_m - \operatorname{sgn}(a_{mk})\beta_k\right]^2 \qquad (3)$$

where $|a_{mk}|$ is the measure of connection intensity between SNP x_m and x_k. The first term of (3) is a summation of MCPs, promoting sparsity in the estimated model. The role of the second term is to encourage smoothness among the coefficient profiles of the related SNPs. Furthermore, the second term can be associated with a Laplacian matrix for a properly defined undirected weighted graph corresponding to the SNPs. As shown in Huang et al. [6], the penalty in (3) is capable of taking correlation structure into account without introducing extra bias, consequently it outperforms a large class of network–constrained penalty functions. The oracle property has also been rigorously established. Therefore, we choose (3) and extend it to the penalized logistic regression model for the analysis of case control type 2 diabetes data.

The network adjacency measure, $|a_{mk}|$, is perhaps the most crucial characteristic in a network to quantify strength of connection between any two nodes (Zhang and Horvath [9]). Denote $A = (a_{mk},\ 1 \le m,\ k \le p)$ as the adjacency matrix, and let r_{mk} be the corresponding Pearson correlation coefficient. We propose to use $a_{mk} = r_{mk}^{\alpha} \cdot I\{\ |\ r_{mk}| > r_c\}$ with α =5. This measure keeps the strong correlations while downweighing the weak ones. In addition, it guarantees that a_{mk} and r_{mk} have the same sign. Compared with the threshold r_c which determines whether the edge joins the corresponding nodes in a network, the power only denotes the relative strength of connection, and does not influence the network structure. Thus α can be chosen via an ad hoc fashion. The correlation cutoff r_c is calculated based on the Fisher transformation $z_{mk} = 0.5\log((1 + r_{mk})/(1 - r_{mk}))$. If the correlation between m^{th} and k^{th} predictor is zero, then $\sqrt{n-3}\, z_{mk}$ approximately follows a standard normal distribution $N(0,1)$, which can be used to determine a threshold c for $\sqrt{n-3}\, z_{mk}$. Subsequently, the corresponding threshold for r_{mk} is $r_c = \frac{\exp\left(2c/\sqrt{n-3}\right)-1}{\exp\left(2c/\sqrt{n-3}\right)+1}$. Such a network is weighted and sparse. We acknowledge that there are other ways of constructing the network adjacency matrix, and conjecture that they are equally applicable. Since our main purpose is not to compare the constructions of different networks, we focus on this particular network structure in this paper.

Computation

Huang et al. [6] adopted a coordinate descent algorithm to obtain the sparse Laplacian shrinkage estimate when the continuous response variable follows a normal distribution. However, this cannot be applied to a binary response directly. We develop a coordinate descent based iteratively reweighed least squares (IRLS) algorithm for the logistic regression, which yields a form the same as the quadratic approximation to the penalized objective function based on Taylor expansion about current estimates. Denote $\beta^{(d)}$ as the value of the regression coefficients at the beginning of the dth iteration, the quadratic approximation to (2) is

$$Q(\beta) \approx -\frac{1}{2n}(\tilde{y}-X\beta)^T W(\tilde{y}-X\beta) + P(\beta;\ \lambda,\ \gamma)$$

where W is an $n \times n$ diagonal matrix of weights with elements $w_i = \pi_i(1 - \pi_i)$, and $\tilde{y}$ is the working response, defined as

$$\tilde{y} = X\beta^{(d)} + W^{-1}(y-\pi)$$

where $\pi = (\pi_1, \ldots, \pi_n)^T$ is evaluated at current parameters $\beta^{(d)}$. The residuals after each iteration can be expressed as

$$r = \tilde{y}-X\beta^{(d)} = W^{-1}(y-\pi)$$

Let $v_m = n^{-1}X_m^T W X_m$. For a standardized design matrix X, we can have

$$z_m = n^{-1}X_m^T W\left(\tilde{y}-X_{-m}\beta_{-m}\right) = v_m\left(n^{-1}X_m^T r + \beta_m^{(d)}\right)$$

Here, v_m needs to be re-weighted in every iteration, leading to increased computational cost. As the Hessian terms can be approximated by an exact upper bound

Table 1 Simulation for SNP data: mean(sd) of true positives (TP) and false positives (FP) based on 100 replicates. Upper panel: $(n, p) = (500, 750)$; Lower panel: $(n, p) = (1000, 1500)$

		A1	A2	A3	A4
AR	TP	41.04(7.07)	39.28(7.40)	61.14(2.86)	53.30(4.23)
$\rho = 0.1$	FP	37.02(28.65)	32.02(27.49)	169.38(18.79)	83.58(21.61)
AR	TP	64.86(6.89)	41.46(8.62)	64.66(2.89)	57.22(3.96)
$\rho = 0.5$	FP	32.12(30.54)	21.84(24.23)	136.22(11.93)	60.10(11.60)
AR	TP	74.98(0.14)	31.22(4.57)	64.18(2.68)	51.98(3.11)
$\rho = 0.9$	FP	9.74(13.31)	4.30(3.90)	88.84(12.52)	30.52(9.67)
Block	TP	47.70(8.62)	45.82(8.92)	62.70(2.75)	55.70(4.20)
$\rho = 0.1$	FP	61.40(77.21)	52.22(71.40)	162.66(18.66)	79.52(19.80)
Block	TP	67.92(6.35)	39.62(7.05)	65.30(3.23)	57.00(3.68)
$\rho = 0.5$	FP	31.40(22.18)	15.74(13.65)	116.16(13.10)	43.96(13.36)
Block	TP	72.06(4.16)	30.28(6.08)	64.08(2.29)	50.94(3.44)
$\rho = 0.9$	FP	12.22(16.30)	5.08(6.22)	83.30(12.76)	27.86(10.81)
AR	TP	94.20(12.15)	94.16(12.28)	126.22(4.55)	113.78(6.27)
$\rho = 0.1$	FP	77.10(50.90)	77.62(51.35)	317.70(22.91)	165.26(31.10)
AR	TP	142.68(3.61)	84.28(10.71)	130.12(3.79)	116.66(4.48)
$\rho = 0.5$	FP	28.12(25.50)	30.82(26.16)	233.08(16.09)	103.42(20.82)
AR	TP	149.96(0.20)	64.40(9.54)	124.06(3.05)	98.96(4.46)
$\rho = 0.9$	FP	33.86(33.25)	9.26(9.25)	105.46(17.51)	25.74(9.43)
Block	TP	101.22(12.37)	98.96(13.55)	132.46(3.82)	121.84(4.57)
$\rho = 0.1$	FP	72.18(45.32)	67.76(46.36)	274.58(18.19)	129.70(19.58)
Block	TP	145.68(5.93)	75.84(9.18)	129.94(3.13)	114.92(4.25)
$\rho = 0.5$	FP	62.78(58.00)	16.22(13.47)	144.26(17.69)	45.00(12.74)
Block	TP	147.40(8.42)	56.32(9.93)	120.78(4.10)	92.62(6.34)
$\rho = 0.9$	FP	27.56(30.66)	6.36(7.59)	81.14(13.58)	19.60(9.02)

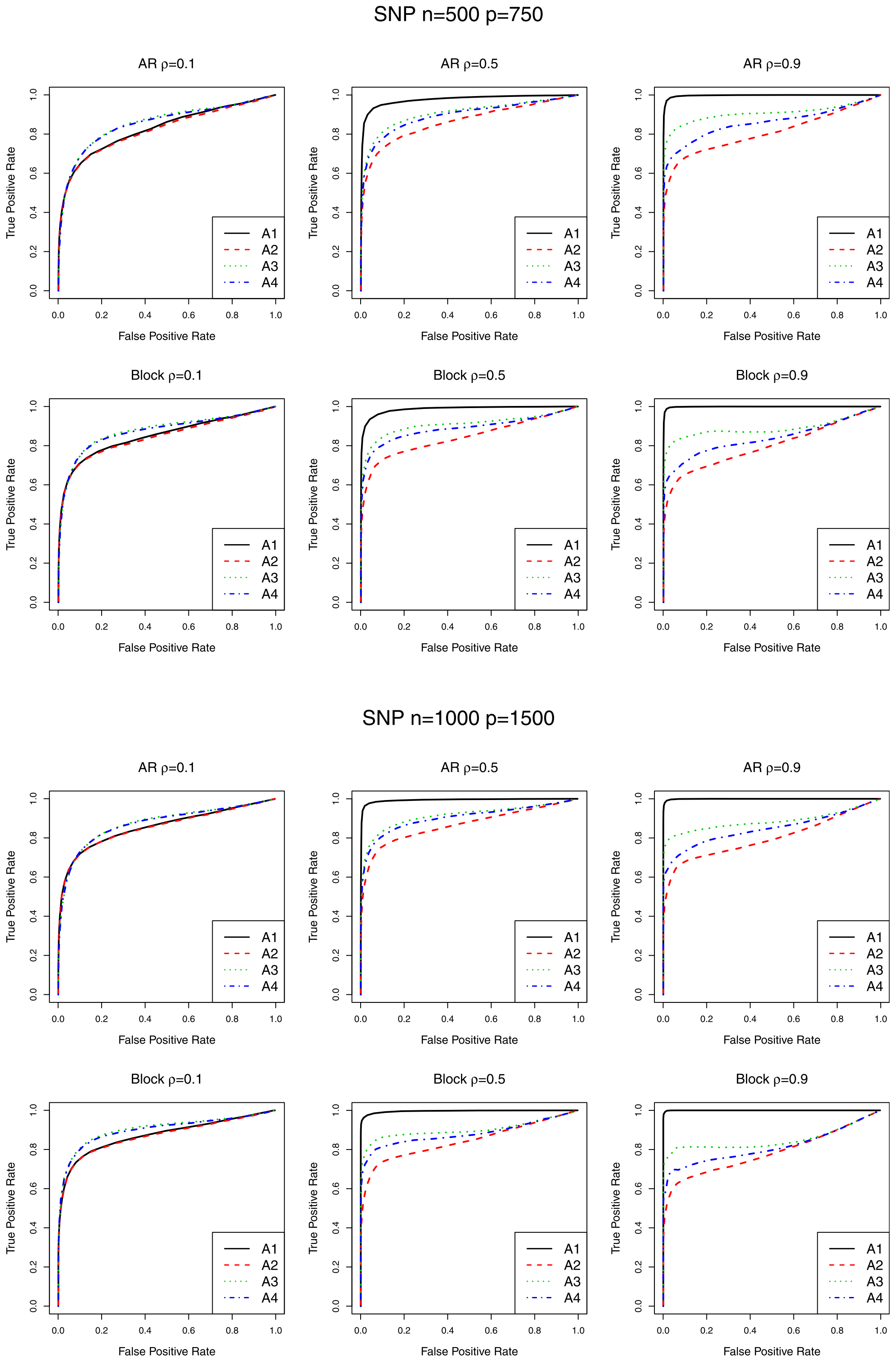

Fig. 1 ROC curves for SNP data. ROC curves corresponding to Table 1

(Krishnapuram et al. [10]), we can set w_i all equal to $\frac{1}{4}$. Define u_m and t_m at iteration d as

$$\begin{aligned} u_m &= z_m + \lambda_2 \sum_{k=m+1}^{p} |a_{mk}|\beta_k \\ t_m &= \frac{1}{4} + \lambda_2 \sum_{k=m+1}^{p} |a_{mk}| \end{aligned} \tag{4}$$

Then the close form update of $\beta^{(d+1)}$ can be obtained as

$$\beta_m^{(d+1)} = \begin{cases} \dfrac{S(u_m,\ \lambda_1)}{t_m - 1/\gamma} & \text{if } |u_m| \le t_m\gamma\lambda_1 \\ \dfrac{u_m}{t_m} & \text{if } |u_m| > t_m\gamma\lambda_1 \end{cases} \tag{5}$$

where $S(\cdot)$ is the soft thresholding function. With fixed tuning parameters λ_1 and λ_2, the coordinate descent algorithm proceeds as follows.

Algorithm Coordinate descent algorithm for penalized network–constrained logistic regression

```
Initialize $d = 0$ and $\beta^{(d)} = 0$ component-wise.
repeat
    $\eta \leftarrow X\beta$
    $\pi \leftarrow e^{\eta} / (1 + e^{\eta})$
    $r^{(m)} \leftarrow (y - \pi) / v$
    for $m = 1, 2, ..., p$
        compute $u_m$ and $t_m$ via (4)
        update $\beta_m^{(d+1)}$ via (5)
        update $r^{(d+1)} \leftarrow r^{(d)} - (\beta_m^{(d+1)} - \beta_m^{(d)}) X_m$
until convergence
```

The convergence is achieved when the L_2 difference between β estimates from two contiguous iterations is smaller than a predefined threshold. Tuning parameters λ_1 and λ_2 control the sparsity in SNP selection and smoothness among coefficient profiles, respectively. They can be chosen from cross validation based methods. We search over a two-dimensional discrete grid of values for λ_1 and λ_2, and select the optimal pair in terms of testing misclassification rate. In penalized logistic regression, regularization parameter γ needs to be larger than $1/w_i$ for MCP. We set it as 4.5 in the simulation study since it has been observed that smaller γ yield slightly better results.

Results

Simulation

We evaluate the performance of the proposed approach through extensive simulation studies. Both categorical and continuous predictors are considered, and they correspond to SNP and gene expression data, respectively. We first generate a $n \times p$ matrix of gene expressions, where $n = 500$ and $p = 750$, from a multivariate normal distribution. For the 750 genes, there are 100 clusters with 5 genes per cluster. The gene expressions have been marginally standardized. We consider two correlation structures. (1) the auto-regression (AR) structure, in which gene i and j within the same cluster have correlation coefficients $\rho^{|i-j|}$, and they are independent cluster–wisely. (2) the block structure, in which the within cluster correlation coefficient is ρ, and gene expressions in different clusters are independent. We consider ρ =0.1, 0.5 and 0.9 for both structures. In addition to the 500 by 750 matrix of gene expressions, a 1000 by 1500 matrix has also been generated with 150 clusters and 10 genes per cluster following the same correlation structures. The SNP data are simulated by dichotomizing expression values of each gene at the 1st and 3rd quartiles, with the 3–level (2,1,0) for genotypes (AA,Aa,aa) respectively. For both combinations of (n, p), (500, 750) and (1000, 1500), 10% of clusters are randomly selected to have nonzero regression coefficients, which are generated from *Unif* [0.25,0.75]. The binary response can subsequently be simulated. We choose the tuning parameters based upon the prediction

Table 2 Simulation for Gene expression data: mean(sd) of true positives (TP) and false positives (FP) based on 100 replicates. Upper panel: (n, p) = (500, 750); Lower panel: (n, p) = (1000,1500)

		A1	A2	A3	A4
AR	TP	43.50(8.64)	40.48(8.48)	61.46(2.92)	53.50(4.59)
$\rho = 0.1$	FP	49.50(45.54)	35.58(32.92)	163.08(15.56)	76.92(19.67)
AR	TP	68.74(9.23)	38.56(7.04)	64.46(2.56)	55.54(3.69)
$\rho = 0.5$	FP	29.64(25.13)	17.06(15.23)	127.36(17.11)	54.94(16.08)
AR	TP	74.34(2.00)	27.68(5.58)	65.30(1.62)	45.50(3.11)
$\rho = 0.9$	FP	10.48(13.50)	3.50(3.72)	76.82(14.10)	23.80(9.96)
Block	TP	44.92(8.75)	42.92(7.96)	64.20(2.91)	56.82(3.70)
$\rho = 0.1$	FP	40.58(40.79)	30.84(24.45)	161.44(13.32)	77.24(17.99)
Block	TP	72.72(4.01)	38.94(6.86)	65.36(2.84)	56.88(3.54)
$\rho = 0.5$	FP	22.08(30.01)	15.06(18.84)	107.18(12.58)	38.70(11.31)
Block	TP	70.12(4.29)	25.24(4.38)	64.62(2.58)	43.50(3.22)
$\rho = 0.9$	FP	5.88(10.48)	1.92(1.52)	75.42(11.49)	23.16(7.77)
AR	TP	88.86(15.09)	86.72(15.32)	126.16(4.51)	113.38(5.89)
$\rho = 0.1$	FP	69.58(56.10)	61.56(49.57)	312.28(24.32)	159.98(27.91)
AR	TP	146.14(2.65)	81.68(12.44)	129.88(3.27)	115.14(4.93)
$\rho = 0.5$	FP	43.62(37.20)	24.20(20.15)	217.86(17.82)	93.60(16.22)
AR	TP	149.42(2.26)	52.64(7.78)	122.50(4.03)	82.42(5.24)
$\rho = 0.9$	FP	27.98(37.43)	4.22(5.28)	97.06(14.09)	23.26(8.58)
Block	TP	91.70(12.04)	88.92(12.70)	131.92(3.26)	120.76(4.32)
$\rho = 0.1$	FP	47.36(31.01)	41.96(28.95)	264.58(20.20)	124.84(17.73)
Block	TP	148.38(4.95)	74.02(10.59)	127.70(3.70)	110.94(4.60)
$\rho = 0.5$	FP	24.78(29.56)	17.46(12.93)	127.80(17.06)	37.10(12.48)
Block	TP	145.10(4.85)	45.60(7.37)	117.90(3.71)	73.56(4.73)
$\rho = 0.9$	FP	18.54(32.97)	2.74(2.47)	69.06(12.25)	14.06(5.94)

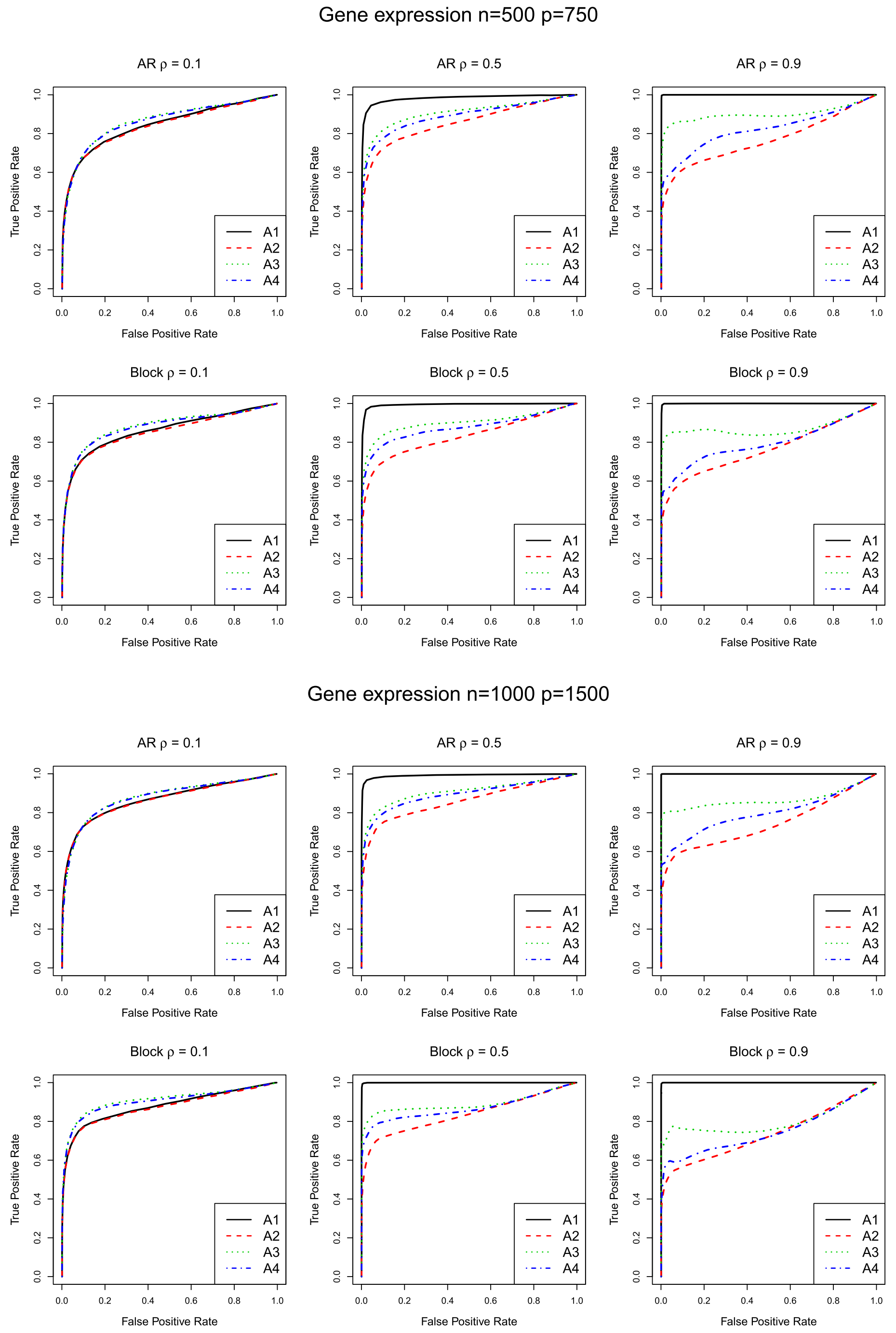

Fig. 2 ROC curves for Gene Expression data. ROC curves corresponding to Table 2

performance of the corresponding model in an independently simulated validation dataset.

For comparison, we consider three alternative approaches, LASSO, elastic net and MCP. LASSO is perhaps so far the the most widely used penalization approach for the analysis of genomic data. In contrast to LASSO, elastic net encourages the grouping effects among genomic features. MCP is equivalent to the proposed approach when $\lambda_2 = 0$ in (3). Comparison with MCP as well as elastic net will directly demonstrate the advantage of each penalty term in the formulation (3). For convenience, we term the network approach, MCP, elastic net and LASSO as A1, A2, A3 and A4, respectively.

Simulation results for the SNP data are tabulated in Table 1. We can observe that from the upper panel where $(n, p) = (500, 750)$, A1 (network) and A2 (MCP) have similar performance when correlation is low. As correlation increases, the proposed one starts to outperform A2. For example, when $\rho = 0.9$ under AR correlation structure, A1 can identify most of the 75 true positives, 74.98 (sd 0.14), with a small number of false positives 9.74 (sd 13.31). A2 identifies similar number of false positives with a much lower number of true positives 31.22 (sd 4.57). Out of all the four approaches, A3 (elastic net) always has the largest false positives and A4 (LASSO) is inferior to A2 in general. Consistent patterns have been observed under other scenarios in Table 1. In addition, we examine the performances using the ROC curves. The ROC curves corresponding to Table 1 are given in Fig. 1, which clearly show that A1 outperforms A2–4. Additional simulation results for gene expression data are given in Table 2 and Fig. 2, which also demonstrate the merit of the network approach over the alternatives when moderate to strong correlation exists among genetic variants. To further examine the performance of the proposed approach, we also conduct simulation under $n = 500$ and $p = 1500$. Results are summarized in Table 3 and Fig. 3. Both the identification accuracy in Table 3 and ROC curves in Fig. 3 demonstrate the superiority of the proposed A1 over alternatives.

In addition to the identification results and the ROC curves, we acknowledge that plots of piece-wise solution path can be adopted to investigate the similarity and difference among different regularization methods, especially when the number of predictors (features) entering the model is moderate or small. In our simulation study, the number of important SNPs and gene expressions is large, therefore such an approach is not pursued.

Table 3 Simulation for $(n, p) = (500, 1500)$: mean(sd) of true positives (TP) and false positives (FP) based on 100 replicates. Upper panel: SNP data; Lower panel: gene expression data

		A1	A2	A3	A4
AR	TP	48.52(12.31)	37.57(12.59)	57.55(14.03)	46.60(14.58)
$\rho = 0.1$	FP	78.29(41.61)	64.29(51.34)	95.11(36.75)	63.32(30.63)
AR	TP	126.91(10.98)	59.37(10.48)	92.57(6.94)	78.83(6.70)
$\rho = 0.5$	FP	78.69(33.05)	54.19(29.45)	139.60(27.36)	87.37(17.78)
AR	TP	148.43(9.83)	47.82(12.45)	105.29(5.34)	80.05(4.81)
$\rho = 0.9$	FP	61.36(48.01)	21.41(60.62)	135.23(20.22)	80.03(12.34)
Block	TP	67.16(12.67)	52.23(11.89)	81.51(7.92)	70.51(7.25)
$\rho = 0.1$	FP	91.44(40.23)	63.03(42.99)	124.56(28.34)	82.16(19.66)
Block	TP	146.63(7.27)	57.24(13.70)	105.11(4.96)	89.64(4.46)
$\rho = 0.5$	FP	81.27(63.42)	32.85(42.69)	143.77(19.63)	87.04(12.76)
Block	TP	143.11(5.33)	43.92(10.42)	105.69(5.16)	82.12(5.11)
$\rho = 0.9$	FP	28.85(41.83)	11.99(19.14)	133.15(20.07)	75.53(12.27)
AR	TP	47.41(10.56)	45.27(11.08)	61.12(12.18)	49.37(13.06)
$\rho = 0.1$	FP	80.08(38.04)	73.56(39.11)	100.51(34.81)	64.09(28.80)
AR	TP	137.85(9.22)	50.61(9.79)	96.19(6.07)	81.23(5.95)
$\rho = 0.5$	FP	74.79(30.71)	31.28(14.74)	142.99(25.91)	91.92(16.12)
AR	TP	148.80(3.65)	38.57(9.09)	107.95(4.73)	70.91(4.23)
$\rho = 0.9$	FP	40.33(31.15)	7.09(10.13)	129.96(17.41)	77.48(11.09)
Block	TP	60.37(11.61)	53.43(12.30)	86.67(6.85)	74.79(5.96)
$\rho = 0.1$	FP	83.04(48.75)	60.64(41.84)	133.73(26.74)	88.28(18.60)
Block	TP	140.67(14.14)	53.89(13.26)	104.83(4.67)	86.65(4.25)
$\rho = 0.5$	FP	73.60(66.06)	27.35(31.89)	138.88(17.85)	82.63(13.30)
Block	TP	145.12(5.00)	33.07(7.61)	106.67(5.81)	64.76(4.71)
$\rho = 0.9$	FP	19.43(23.30)	3.79(6.54)	123.37(19.61)	67.60(9.78)

Real data analysis

As described in the background section, we analyze Nurses' Health Study (NHS), a case control study of type 2 diabetes which are part of the Gene Environment Association Studies. Detailed information of the datasets are available from Hu et al. [11]. We focus on SNPs in several important pathways potentially related to T2D. They are the Wnt signaling pathway, cell cycle pathway and p53 signaling pathway. After cleaning the data through matching phenotypes and genotypes, removing SNPs with minor allele frequency (MAF) less than 0.05 or deviation from Hardy–Weinberg equilibrium, the working dataset contains 3391 subjects. There are 5079 SNPs in the Wnt signaling pathway, and 3793 SNPs in the cell cycle and p53 signaling pathway.

We first apply all the 4 methods on the Wnt signaling pathway. The A1–4 identify 834, 841,847 and 848 SNPs that are associated with T2D, respectively. As a representative example, we examine closely gene DAMM1 and its subnetwork. DAMM1 is reported to be associated with diabetic nephropathy, a common complication of type 2 diabetes (Sapienza et al. [12] and Kavanagh et al. [13]). The upper panel of Fig. 4 shows the subnetwork of DAMM1, where the red nodes indicate the SNPs from DAMM1. In the subnetworks, thickness of edges denotes the strength of correlation between SNPs. It can be clearly observed that the proposed approach has identified much more highly correlated SNPs, since the interconnections among SNPs have been accommodated by the approach that incorporates the network structure information. The network approach (A1) selects 19 SNPs and 15 belong to DAMM1, while other 3 approaches only identify 9,6 and 6 SNPs correspondingly. A1 leads to a more tightly connected network, which is consistent with our findings in the simulation study that it promotes the interconnections among SNPs. Furthermore, the proposed one identifies SNP rs1252906, which plays a crucial role in the progression of nephropathy (Kavanagh et al. [13]). Other methods fail to identify this important SNP. The genes corresponding to the SNPs in the subnetworks are given in the upper panel of Fig. 5.

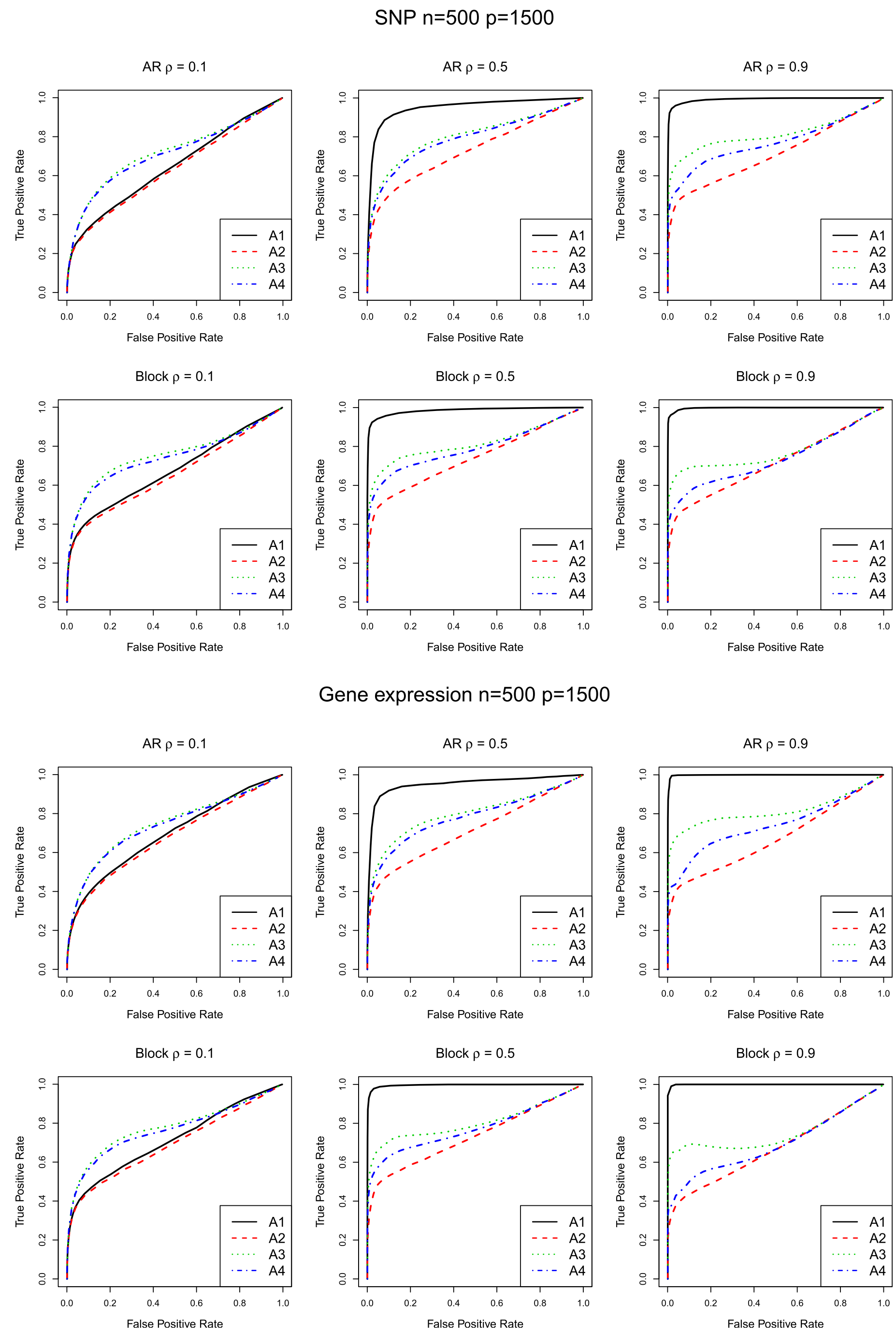

Fig. 3 ROC curves for *(n, p)* = (500,1500). ROC curves corresponding to Table 3

The analysis has also been carried out on the SNPs combined from the cell cycle pathway and p53 signaling pathway. There are 814, 828, 827 and 829 SNPs identified by A1–4 correspondingly. We focus on the subnetwork of gene CASP9, which is one of the key players in inducing cell apoptosis (Cnop et al. [14]). Previous studies show that CASP9 is associated with diabetic retinopathy, a common and serious complication of type 2 diabetes (Baharian et al. [15] and Looker et al. [16]). In the NHS study, CASP9 has a total of 11 SNPs. The subnetwork of CASP9 is shown in the lower panel of Fig. 4. The proposed method identifies a subnetwork which has 7 SNPs from CASP9, and the other 7 SNPs from gene CELA2A, CELA2B and DNAJC16. Both CELA2A and CELA2B encode protein elastases, which hydrolyze elastin and many other proteins. DNAJC16 is a member of

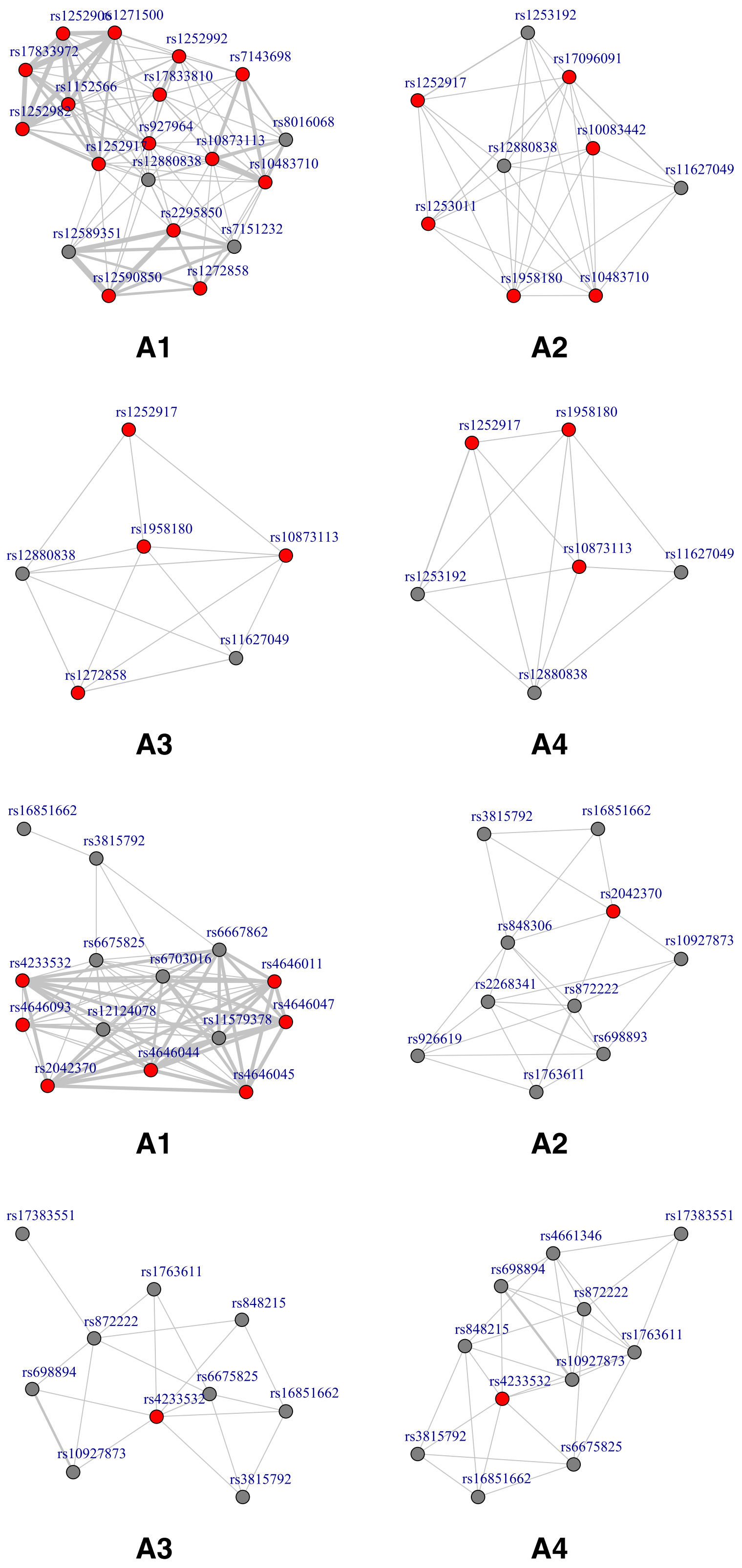

Fig. 4 Subnetworks of DAAM1 (upper panel) and CASP9 (lower panel). SNPs connected in the network are joined with edges

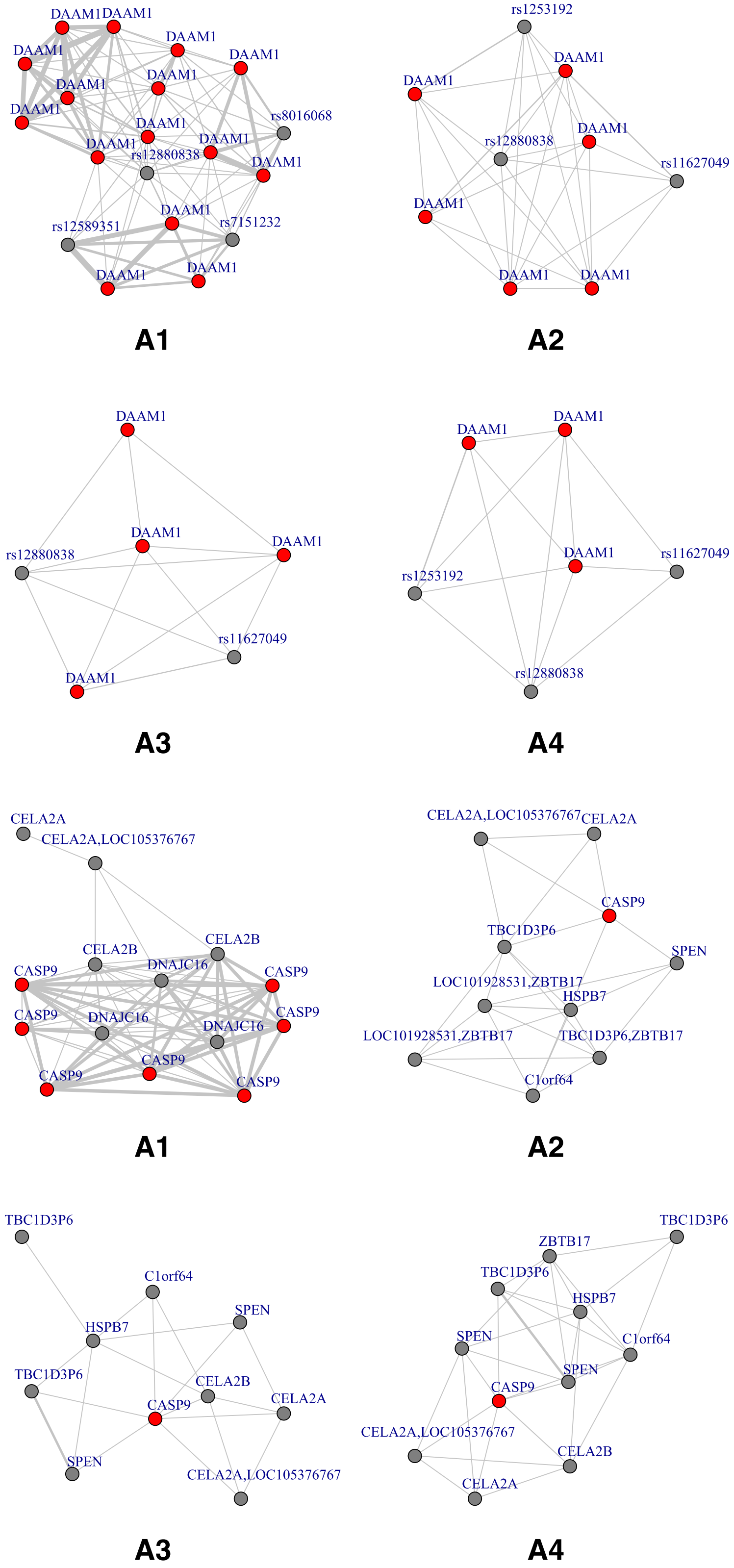

Fig. 5 Subnetworks of DAAM1 (upper panel) and CASP9 (lower panel). Gene names are corresponding to the SNP id in Fig. 4

heat shock protein family (Hsp40). And it has been found in multiple studies that Hsp40 are related to cell apoptosis in type 2 diabetes (Laybutt et al. [17] and Chien et al. [18]. It is very interesting that CELA2A, CELA2B, CASP9 and DNAJC16 locate on chromosome 1 as a cluster. CELA2A is also identified by the rest 3 approaches, while CELA2B is only not in the subnetwork identified by A2(MCP). Overall, the network effects of CASP9, CELA2A, CELA2B and DANJC16 on type 2 diabetes, especially diabetic retinopathy, worth further investigations.

Discussion

In this paper, we develop a network–based regularized logistic regression model for the analysis of high dimensional genetic data and identification of important SNPs in the case–control study of type 2 diabetes. Advancing from existing studies, the proposed one has desired property to take correlation pattern among genetic variants into account without incurring extra bias. We provide an efficient iteratively reweighed least squares (IRLS) algorithm within the coordinate descent framework. The computational cost has been significantly reduced due to convenient approximations to the original regularized log likelihood function. Simulation demonstrates that the proposed one outperforms closely related alternatives.

Computation feasibility is an important practical consideration for high–dimensional regularization methods. In simulation, the CPU time (in minutes) of applying the proposed method on 100 replicates of simulated SNP data with $n = 1000$, $p = 1500$ and AR structure is 218.3 on a regular laptop. In the case study, the CPU time for analyzing the Wnt pathway with $n = 3391$ and $p = 5079$ is 70.16. The proposed method can be potentially applied to larger datasets with a reasonable computation time. It has been widely acknowledged in Fan and Lv [19], Jiang et al. [20] and studies alike that regularization methods have to be coupled with screening strategy to accommodate ultra-high dimensional data from for instance, large-scale GWAS studies. The proposed network constrained regularization method can be implemented in such a two stage framework. Further investigations are intriguing but beyond the scope of this paper, and will be postponed to the future.

The methodological development in this article has been partially motivated by the analysis of the datasets from Nurese's Health Study (NHS). In the past, this study has been extensively investigated under marginal methods ([21] and [22]) which ignores the joint effects of the SNPs. In addition, although studies like Wu, Cui and Ma [23] consider the effects of SNPs jointly within the penalization framework for continuous phenotypes, the correlation among the SNPs still have not been fully taken into account. The proposed approach quantifies the strength of correlation among SNPs via network structure and is able to incorporate the correlation in the identification of important SNPs. In the case study, we have identified interesting subnetworks with respect to genes closely related to T2D. In this work, we have focused on methodological development. More thorough bioinformatics and functional investigations will be needed in the future to fully understand the identified results.

Conclusions

The network-constained logistic regulaization method proposed in this study has demonstrated superior performance in identifying important genetic variants from both simulation study and the Nurese's Health Study, a case–control study of type 2 diabetes with high dimensional SNP measurements. The network term in the regularized loss function accomodates the LD widely present among SNPs, which guarantees the advantage of the developed one over the competing alternative methods.

Abbreviations

AR: Auto-regression; GWAS: Genome–wide association studies; IRLS: Iteratively reweighed least squares; LASSO: Least Absolute Shrinkage and Selection Operator; LD: Linkage disequilibrium; MAF: Minor allele frequency; MCP: Minimax concave penalty; NHS: Nurses's Health Studies; SLS: Sparse Laplacian shrinkage; T2D: Type 2 diabetes

Acknowledgements

The dataset used for the analyses described in this manuscript has been obtained from dbGaP through accession number phs000091.v2.p1.

Funding

Funding support for the GWAS of Gene and Environment Initiatives in Type 2 Diabetes was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004399). This study was partially supported by an award from the Faculty Enhancement Program, Kansas State University.

Availability of data and materials

Authorized access should be granted before accessing the data. Therefore we are not authorized to deposit the data in publicly available repositories, or to present the data within the manuscript and/or additional supporting files. Applications to access the data should be sent to dbGap (accession number phs000091.v2.p1). For more information, please refer to NIH dbGap (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1).

Authors' contributions

JR developed the model, performed the statistical analysis on both simulated and real data, and drafted the manuscript. TH, YJ and CW participated in real data analysis. YL provided support in model design. YL, SL and YD provided constructive comments and suggestions. CW conceived the idea, developed the model and finalized the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

This work is a secondary data analysis. The dataset has been applied through NIH dbGap (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1). In the dataset, the patient information has been de-identified. As indicated from the dbGap website under section *Authorized Access/Use Restrictions*, IRB is not required for accessing and using the data. According to the original publication, "The study was approved by the institutional review board of Brigham and Women's Hospital in Boston; completion of the self-administered questionnaire was considered to imply informed consent." For more information regarding study population, please refer to the original publication: Hu, F.B., Manson, J.E., Stampfer, M.J., Colditz, G., Liu, S., Solomon, C.G., Willett, W.C. (2001) Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N Engl J Med.* 345(11):790–797. PubMed PMID: 11556298.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Statistics, Kansas State University, 1116 Mid-Campus Drive N., 66506 Manhattan, KS, USA. [2]Department of Mathematics, San Francisco State University, San Francisco, CA, USA. [3]Department of Biostatistics, Yale University, New Haven, CT, USA. [4]Division of Nephrology, School of Medicine, Stanford University, Palo Alto, CA, USA. [5]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, USA.

Received: 5 November 2016 Accepted: 25 March 2017
Published online: 16 May 2017

References

1. Wild SH, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care. 2004; 27(10):2569.
2. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, Strawbridge RJ, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 2012;44(9):981–90.
3. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B. 1996;58(1):267–88.
4. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67(2):301–20.
5. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008;24(9):1175–82.
6. Huang J, Ma S, Li H, Zhang CH. The sparse laplacian shrinkage estimator for high-dimensional regression. Ann Stat. 2011;39(4):2021–46.
7. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.
8. Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, et al. The gene, environment association studies consortium (geneva): maximizing the knowledge obtained from gwas by collaboration across studies of multiple conditions. Genet Epidemiol. 2010;34(4):364–72.
9. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4(1):1128.
10. Krishnapuram B, Carin L, Figueiredo MAT, Hartemink AJ. Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans Pattern Anal Mach Intell. 2005;27(6):957–68.
11. Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, Willett WC. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med. 2001;345(11):790–7.
12. Sapienza C, Lee J, Powell J, Erinle O, Yafai F, Reichert J, Siraj ES, Madaio M. DNA methylation profiling identifies epigenetic differences between diabetes patients with ESRD and diabetes patients without nephropathy. Epigenetics. 2011;6(1):20–8.
13. Kavanagh DH, Savage DA, Patterson CC, McKnight AJ, Crean JK, Maxwell AP, McKay GJ. Haplotype association analysis of genes within the wnt signalling pathways in diabetic nephropathy. BMC Nephrol. 2013;14(1):126.
14. Cnop M, Welsh N, Jonas J-C, J¨orns A, Lenzen S, Eizirik DL. Mechanisms of pancreatic *β*–cell death in type 1 and type 2 diabetes many differences, few similarities. Diabetes. 2005;54 suppl 2:97–107.
15. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, et al. The great migration and african-american genomic diversity. PLoS Genet. 2016;12(5):1006059.
16. Looker HC, Nelson RG, Chew E, Klein R, Klein BEK, Knowler WC, Hanson RL. Genome-wide linkage analyses to identify loci for diabetic retinopathy. Diabetes. 2007;56(4):1160–6.
17. Laybutt DR, Preston AM, Akerfeldt MC, Kench JG, Busch AK, Biankin AV, Biden TJ. Endoplasmic reticulum stress contributes to beta cell apoptosis in type 2 diabetes. Diabetologia. 2007;50(4):752–63.
18. Chien V, Aitken JF, Zhang S, Buchanan CM, Hickey A, Brittain T, Cooper GJS, Loomes KM. The chaperone proteins hsp70, hsp40/dnaj and grp78/bip suppress misfolding and formation of *β*–sheet–containing aggregates by human amylin: a potential role for defective chaperone biology in type 2 diabetes. Biochem J. 2010;432(1):113–21.
19. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B (Statistical Methodology). 2008;70(5):849–911.
20. Jiang L, Liu J, Zhu X, Ye M, Sun L, Lacaze X, Wu R. 2higwas: a unifying high-dimensional platform to infer the global genetic architecture of trait development. Brief Bioinform. 2015;16(6):905–11.
21. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome–wide association studies. Am J Epidemiol. 2009;169(2):219–26.
22. Wu C, Cui Y. A novel method for identifying nonlinear gene–environment interactions in case–control association studies. Hum Genet. 2013;132(12): 1413–25.
23. Wu C, Cui Y, Ma S. Integrative analysis of gene–environment interactions under a multi-response partially linear varying coefficient model. Stat Med. 2014;33(28):4988–98.