Published in final edited form as:
Stat Med. 2013 May 30; 32(12): 2127–2139. doi:10.1002/sim.5694.

Network-based Regularization for Matched Case-Control Analysis of High-dimensional DNA Methylation Data

Hokeun Sun and **Shuang Wang**[*,†]
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

Abstract

The matched case-control designs are commonly used to control for potential confounding factors in genetic epidemiology studies especially epigenetic studies with DNA methylation. Compared with unmatched case-control studies with high-dimensional genomic or epigenetic data, there have been few variable selection methods for matched sets. In an earlier article, we proposed the penalized logistic regression model for the analysis of unmatched DNA methylation data using a network-based penalty. However, for popularly applied matched designs in epigenetic studies that compare DNA methylation between tumor and adjacent non-tumor tissues or between pre-treatment and post-treatment conditions, applying ordinary logistic regression ignoring matching is known to bring serious bias in estimation. In this article, we developed a penalized conditional logistic model using the network-based penalty that encourages a grouping-effect of 1) linked CpG sites within a gene or 2) linked genes within a genetic pathway for analysis of matched DNA methylation data. In our simulation studies, we demonstrated the superiority of using conditional logistic model over unconditional logistic model in high-dimensional variable selection problems for matched case-control data. We further investigated the benefits of utilizing biological group or graph information for matched case-control data. The proposed method was applied to a genome-wide DNA methylation study on hepatocellular carcinoma (HCC) where DNA methylation levels of tumor and adjacent non-tumor tissues from HCC patients were investigated using the Illumina Infinium HumanMethylation27 Beadchip. Several new CpG sites and genes known to be related to HCC were identified but were missed by the standard method in the original paper.

Keywords

DNA methylation; Genetic pathways; Matched case-control; Network-based regularization; Penalized conditional logistic; Variable selection

1. Introduction

The matched case-control designs are widely used in genetic epidemiology studies to control for potential confounding variables such as gender, age, etc. Epigenetic studies, the studies

[*]Correspondence to: Shuang Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA.
[†]sw2206@columbia.edu

of inheritable changes caused by nonsequence-based alterations that are inherited, have been a new recent focus in genetic studies. DNA methylation, among many other epigenetic changes, is a molecular modification of DNA that is crucial for normal development. It has been very common for DNA methylation studies to compare DNA methylation levels of tumor and adjacent non-tumor tissues or to compare DNA methylation levels before and after a treatment/an intervention because too many factors are known to greatly affect DNA methylation levels.

Our motivating example is a recent study of Shen et al. [1], where DNA methylation levels of tumor and adjacent non-tumor tissues from 62 Taiwanese hepatocellular cancer (HCC) cases were investigated to identify differentially methylated genes between tumor and adjacent non-tumor tissues using the Illumina Infinium HumanMethylation27 Beadchip. In the original paper, the paired t-test was conducted at each CpG (Cytosine-phosphate-Guanine) site across 26,486 genome-wide autosomal CpG sites. A Bonferroni adjustment, although known to be extremely conservative, was employed to provide a list of significant CpG sites that are differentially methylated between tumor and adjacent non-tumor tissues for HCC patients. We note that testing each CpG site one at a time does not take into account the correlations existed among methylation levels of multiple CpG sites within a gene [2].

In an earlier article, we developed the penalized logistic regression model for the analysis of correlated DNA methylation data using a network-based penalty and identified important CpG sites potentially associated with ovarian cancer [2]. However, this method was designed for unmatched case-control high-dimensional data. Levin and Paik [3] have shown that standard unconditional logistic regression analysis for matched case-control data ignoring matching may lead to biased log odds ratio estimates. Hansson and Khamis [4] investigated the performance of conditional and unconditional maximum likelihood estimation on matched data with different rates of missing observations and recommended to use the conditional logistic regression when matching ratio and sample sizes are relatively small. In fact, conditional analysis should be performed for matched case-control data, whenever possible.

However, there has been not only little discussion of selection performance of unconditional logistic regression analysis for matched high-dimensional data ignoring matching but also few variable selection methods specifically designed for matched case-control high-dimensional data. In this article, we developed the penalized conditional logistic regression model for matched case-control high-dimensional data where we employed the network-based penalty to either utilize biological group information such as correlated DNA methylation levels of multiple CpG sites within a gene or to utilize graph information such as graphically constrained gene expression levels (or gene-level DNA methylation levels) within a genetic pathway. Through extensive simulation studies, we demonstrated the superiority of conditional logistic analysis on matched case-control data for high-dimensional variable selection problems over unconditional logistic analysis. We further demonstrated the advantages of incorporating prior biological information into matched case-control genetic association studies. Chen et al. [5] have recently showed that linked genes within a genetic pathway are likely to have the same association signals with a disease

from a genome-wide association studies (GWAS) of Crohn's disease. Although there have been several noticeable attempts to incorporate prior knowledge of biological network into GWAS [5–8], all are based on linear regression models with gene expression data, not for matched case-control designs.

We applied the proposed penalized conditional logistic regression model to the genome-wide DNA methylation study on HCC where DNA methylation levels of tumor and adjacent non-tumor tissues from HCC patients were investigated using the Illumina Infinium HumanMethylation27 Beadchip. Several new CpG sites and genes that are known to be related to HCC were identified but were missed by the standard method in the original paper.

2. Methods

Let us denote the dataset of the i-th individual by (y_i, x_i, δ_i); $i = 1; \ldots, n$, where $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is the p dimensional DNA methylation data, and $y_i = 1$ for the case and $y_i = 0$ for the matched control, and δ_i indicates the stratum of the i-th individual, $\delta_i \in \{1,\ldots, K\}$. We assume that the k-th stratum has one case and $n_k - 1$ controls, i.e., $\sum_{i=1}^{n} I(\delta_i=k)=n_k$ and $\sum_{i=1}^{n} y_i I(\delta_i=k)=1$, where $I(\cdot)$ is an indicator function. For example, $n_k = 2$ for the 1: 1 matched design, and $n_k = m + 1$ for the 1: m matched set. Following Fleiss et al. [9], the conditional logistic likelihood given by $n_k \geq 2$ for all $k = 1, \ldots, K$ strata can be written as

$$L(\boldsymbol{\beta}|n_1,\ldots,n_K)=\prod_{k=1}^{K}\frac{\exp(\sum_{j\in\Delta_k} x_j^T\boldsymbol{\beta}y_j)}{\sum_{j\in\Delta_k}\exp(x_j^T\boldsymbol{\beta})}, \quad (1)$$

where $\Delta_k = \{j \leq n: \delta_j = k\}$ consists of n_k indices, and $\beta = (\beta_1, \ldots, _p)^{\mathrm{T}}$ is the parameter vector of interest.

It is noticeable that function (1) could be viewed as the partial likelihood of the stratified Cox proportional hazards model if cases are defined as events of death and controls are censoring. Actually, the partial likelihood of the Cox model can be exactly the same as (1) under one particular circumstance when only the matched controls are at risk with one death occurring for each stratum. In other words, both survival and censoring timelines should not be overlapped between strata, and the event of death should always happen earlier than censoring within strata. In contrast to matched case-control studies, considerably various regularization procedures for the Cox proportional hazards model have been developed over the last few years [10–14]. However, these methods except Goeman [13] do not allow stratification of data, and all survival and censoring times have the same starting point. As a result, most of selection methods developed for the Cox model cannot be directly applied to the matched case-control design. Although the l_1 penalized Cox model developed by Goeman [13] can be used as a penalized conditional logistic regression, we have shown that l_1 penalization is likely to ignore the correlations of CpG sites and thus leads to poor selection performance [2]. This motivated us to develop a new variable selection procedure specifically for matched DNA methylation data.

Similarly as Sun and Wang [2], the penalized conditional logistic likelihood with the network-based penalty function can be written as

$$Q_{\lambda,\alpha}(\boldsymbol{\beta}) = -\frac{1}{n} l(\boldsymbol{\beta}) + \lambda\alpha \|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T S^T L S \boldsymbol{\beta}, \quad (2)$$

where $l(\beta)$ is the log likelihood of (1), and $\lambda > 0$ and $0 \leq \alpha \leq 1$ are tuning parameters to control sparsity and smoothness, respectively. The normalized Laplacian matrix $L = \{l_{uv}\}_{p\times p}$ represents a graph structure of predictors when the network information is provided, and is defined as

$$l_{uv} = \begin{cases} 1 & \text{if} \quad u = v \text{ and } d_u \neq 0 \\ -(d_u d_v)^{-\frac{1}{2}} & \text{if} \quad u \text{ and } v \text{ are linked with each other} \\ 0 & \text{otherwise,} \end{cases}$$

where d_u is the total number of variables linked to the u-th predictor. The diagonal matrix $S = \text{diag}(s_1, \ldots, s_p)$, $s_u \in \{-1, 1\}$ in (2) has the signs of regression coefficients on its diagonal entries, which are often preliminarily estimated from ordinary regression for $p < n$, and ridge regression for $p \geq n$. As pointed out by Li and Li [7], the matrix S can accommodate the problem of failure of local smoothness between linked variables, where two adjacent risk factors have opposite effects on a disease when the corresponding regression coefficients have different signs.

The penalized conditional likelihood function (2) can then be explicitly written as

$$Q_{\lambda,\alpha}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{k=1}^{K} \log\left(\sum_{j\in\Delta_k} \exp(x_j^T\boldsymbol{\beta})\right) - \frac{1}{n}\sum_{i=1}^{n} x_i^T\boldsymbol{\beta} y_i + \lambda\alpha\sum_{u=1}^{p}|\beta_u| + \frac{1}{2}\lambda(1-\alpha)\sum_{u=1}^{p}\sum_{u\sim v}\left(\frac{s_u\beta_u}{\sqrt{d_u}} - \frac{s_v\beta_v}{\sqrt{d_v}}\right)^2,$$

where $u \sim v$ indicates the index set of all linked variables to the u-th predictor. The penalty function consists of a combination of the l_1 norm and squared l_2 norm on degree-scaled differences of coefficients between linked variables, and thus induces both sparsity and smoothness with respect to the correlated or linked structure of the regression coefficients. Subsequently, a desirable grouping effect can be reached by specifying links among regression coefficients in the model, where the coefficients of predictors linked on the network can shrink towards each other, allowing them to borrow information from each other.

The network-based regularization procedure is also applicable even when only distinct group information among variables is provided instead of complex network information. In our recent study [2] under the unmatched case-control setting, we employed ring and fully connected network models for correlated variables within a group, assuming only variables in the same group are linked with each other. In high-throughput DNA methylation data, multiple CpG sites from one gene are usually highly correlated. Sun and Wang [2] also

showed real correlations of CpG sites within each gene using ovarian cancer data. Therefore, we could regard each gene as a group and introduce a grouping effect for CpG sites within one gene. It is a challenging work to precisely capture such correlation patterns. To set them to be linked to each other is an approximation that encourages selection effects for highly correlated CpG sites in the network-based regularization procedure. CpG sites from linked genes within a genetic pathway are also correlated when linked genes within the genetic pathway have similar biological functions and are correlated each other. Thus, we can incorporate genetic pathway information into methylation data analysis to better discover outcome-related CpG sites.

Note that although the network-based regularization procedure tends to select linked variables, it performs individual selection for grouped variables. That is, variables not related to the outcome are not selected even if they are linked to outcome-related variables. This is a great advantage over group selection methods. For example, when causal and neutral genetic sites are located in the same group such as genes or haplotype blocks, selecting only causal sites is ideally preferred. But, the group penalized methods [15, 16] force to select entire groups. In DNA methylation data, causal genes might contain neutral CpG sites, and a genetic pathway might consist of both causal and neutral genes. In conclusion, the network-based regularization procedure is more adequate for analysis of high-dimensional DNA methylation data.

In the current matched case-control setting, the estimate $\hat{\beta}$ that minimizes the objective function (2) can be dexterously obtained for fixed λ and α using the cyclic coordinate descent algorithm [2, 17, 18]. First, the log likelihood $l(\boldsymbol{\beta})$ is replaced by a quadratic approximation using the Taylor expansion [14], and then coordinate descent is employed to solve iteratively weighted least squares. Specifically, given the current estimate $\boldsymbol{\beta}^*=(\beta_1^*,\ldots,\beta_p^*)^T$, we have the following closed form solution to each β_u,

$$\hat{\beta}_u=\frac{\operatorname{sign}(T_u)(|T_u|-\lambda\alpha)_+}{n^{-1}\sum_{i=1}^n w_i(\boldsymbol{\beta}^*)x_{iu}^2+\lambda(1-\alpha)}, \quad (3)$$

where

$$T_u=\frac{1}{n}\sum_{i=1}^n w_i(\boldsymbol{\beta}^*)x_{iu}\left(z_i(\boldsymbol{\beta}^*)-\sum_{j\neq u}x_{ij}\beta_j^*\right)+\lambda(1-\alpha)\frac{s_u}{\sqrt{d_u}}\sum_{u\sim v}\frac{s_v\beta_v^*}{\sqrt{d_v}},$$
$$w_i(\boldsymbol{\beta})=r_i(\boldsymbol{\beta})\,(1-r_i(\boldsymbol{\beta})),$$
$$z_i(\boldsymbol{\beta})=x_i^T\boldsymbol{\beta}+\frac{y_i-r_i(\boldsymbol{\beta})}{r_i(\boldsymbol{\beta})\,(1-r_i(\boldsymbol{\beta}))},$$

and

$$r_i(\boldsymbol{\beta})\frac{\exp(x_i^T\boldsymbol{\beta})}{\sum_{j\in\Delta_{\delta_i}}\exp(x_j^T\boldsymbol{\beta}))}.$$

We then summarize our algorithm that obtains the minimizer $\hat{\beta}$ of the objective function (2) as follows:

1. Initialize $\beta^* = 0$;
2. Compute $w_i(\beta^*)$ and $z_i(\beta^*)$ for $i = 1,\ldots, n$;
3. Update $\hat{\beta}_u$ by (3) cyclically for $u = 1,\ldots, p$;
4. Set $\beta^* = (\hat{\beta}_1,\ldots, \hat{\beta}_p)^T$;
5. Repeat steps 2 – 4 until some convergence criterion is met.

Similar to our previous work on unmatched case-control studies [2], we also employed the selection probability method [19] to handle with tuning parameters, where we compared selection probabilities of predictors and selected top ranked variables by the selection probabilities. One difference of computing selection probabilities in the penalized conditional logistic model from that of the penalized ordinary logistic model is random sampling of strata. Details are shown in Appendix.

3. Simulation Studies

In the simulation studies, we considered two different scenarios that utilize two different biological information. In the first scenario, biological group information such as correlated DNA methylation levels of multiple CpG sites within a gene is utilized. In the second scenarios, graph information such as graphically constrained gene expression levels or graphically constrained gene-level overall DNA methylation levels within a genetic pathway is utilized.

We evaluated selection performance of regularization procedures using several criteria. We first computed true positive rate (TPR) and true negative rate (TNR) which are known as sensitivity and specificity, respectively. We then considered the Mathews Correlation Coefficient (MCC) which is generally regarded as a balanced measure between sensitivity and specificity [20]. They are defined as

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},$$

and

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\,(\mathrm{TP}+\mathrm{FN})\,(\mathrm{TN}+\mathrm{FP})\,(\mathrm{TN}+\mathrm{FN})}},$$

where TP and TN stand for true positives and true negatives, respectively, and FP and FN stand for false positives/negatives. These measures are calculated when a specified number of variables (such as selecting top 30 or 50 ranked variables) are selected. We also constructed the receiver operating characteristic curves (ROC) using TPR and TNR, varying the number of variables selected. The area under the ROC curves (AUC) was then evaluated

to summarize overall selection performance. The estimated AUC using selection probabilities is given by

$$\hat{\mathrm{AUC}}=\frac{1}{t_0(p-t_0)}\sum_{u=1}^{t_0}\sum_{v=t_0+1}^{p}\left[I(\hat{p}_u>\hat{p}_v)+\frac{1}{2}I(\hat{p}_u=\hat{p}_v)\right],$$

where $\hat{p_u}$ is the selection probability of the u-th predictor, and the first t_0 variables are assumed to be true signals. Pepe [21] has shown this estimate using predictive probabilities. Note that this is also equivalent to the Wilcoxon or MannWhitney U-statistic.

3.1. Scenario I: grouped variables

In the first type of simulation studies when predictors within a group are correlated, we compared the performance of the proposed regularization procedure of conditional logistic regression model against that of an unconditional logistic model on matched case-control data.

Data was generated based on retrospective case-control studies where we match one case (Y = 1) to one or more controls (Y = 0) on a matching variable Z, and then observe their risk factors X such as DNA methylation CpG sites. We assumed that X follows a conditional scaled normal distribution given Y and Z, e.g.,

$$X|Y=1,Z=z\sim\sqrt{z}N(\boldsymbol{\mu},\sum),\quad\text{and}\quad X|Y=0,Z=z\sim\sqrt{z}N(0,\sum). \quad (4)$$

Note that we simulated multivariate normal data for simplicity even if real DNA methylation β-values lie between 0 and 1. The logit transformation of the β-values into a normal distribution is often employed for analysis of methylation data [22, 23].

Levin and Paik [3] investigated the estimation bias of unconditional logistic analysis for matched data, considered three different types of distribution for the matching variable, namely, a U-shaped, a bell-shaped, and a uniform distribution. We assumed that the matching variable Z follows a beta distribution *Beta*(a; b), and used different values of a and b such that the three types of the distribution covered. We considered the 1:1 matched design in our simulation studies, where the same matching variable Z was used to generate X in the same stratum.

The number of CpG sites and strata were set to p = 1, 000 and K = 50, respectively. Therefore, we have a total of 1,000 CpG sites with 50 matched sets in this design. We assumed that data consists of 100 genes each of which contains 10 CpG sites. We further assumed each gene has a covariance $\Sigma_{uv} = 0.5^{|u-v|}$, $0 \geq |u - v| \leq 10$ to capture correlation within the gene. The mean vector of cases $\mu = (\beta_1,\ldots, \mu_p)^{\mathrm{T}}$ was defined as follows to vary the strength of the true signals:

$$\mu_u = \begin{cases} (\frac{\delta}{\sqrt{1}}, \frac{\delta}{\sqrt{2}}, \ldots, \frac{\delta}{\sqrt{9}}, \frac{\delta}{\sqrt{10}}, \frac{-\delta}{\sqrt{1}}, \frac{-\delta}{\sqrt{2}}, \ldots, \frac{-\delta}{\sqrt{9}}, \frac{-\delta}{\sqrt{10}}), & 1 \le u \le 20\,(1^{\text{st}} \text{ and } 2^{\text{nd}} \text{gene}), \\ (\frac{\delta}{\sqrt{1}}, \ldots, \frac{\delta}{\sqrt{5}}, 0, 0, 0, 0, 0, \frac{-\delta}{\sqrt{1}}, \ldots, \frac{-\delta}{\sqrt{5}}, 0, 0, 0, 0, 0), & 21 \le u \le 40\,(3^{\text{rd}} \text{ and } 4^{\text{th}} \text{gene}), \\ (0, 0, 0, 0, 0, \frac{\delta}{\sqrt{6}}, \ldots, \frac{\delta}{\sqrt{10}}, 0, 0, 0, 0, 0, \frac{-\delta}{\sqrt{6}}, \ldots, \frac{-\delta}{\sqrt{10}}), & 41 \le u \le 60\,(5^{\text{th}} \text{ and } 6^{\text{th}} \text{gene}), \end{cases}$$

and $\mu_u = 0$ for $u > 60$, where δ was fixed at 1.5. All CpG sites in the first two genes were assumed to be causal sites, but only the first half of CpG sites in the third and fourth genes, and only the last half in the fifth and sixth genes were causal sites. The remaining 94 genes were assumed to include only neutral CpG sites. The group network regularization used in our previous work [2], where we utilized biological group information such that all CpG sites within each gene are assumed to be linked with each other, was conducted for both unconditional and conditional logistic models. The selection probabilities of both models were computed based on 150 resamplings.

The selection performance of the penalized logistic (plogit) and conditional logistic (pclogit) model are displayed in Tables 1. The results are shown along with different *Beta* distributions of the matching variable, where both *Beta*(0.3, 0.3) and *Beta*(0.5, 0.5) form U-shaped distributions, while *Beta*(1, 1) and *Beta*(3, 3) form a uniform distribution and a bell-shaped distribution, respectively. We selected top 40, 60, and 80 ranked CpG sites based on their selection probabilities to evaluate TPR, TNR, and MCC. All simulation results were summarized as their averages and standard errors over 100 replications.

It is clear that pclogit overall identified more causal sites than plogit when the same number of CpG sites were selected in both designs. Particularly, pclogit outperforms plogit when matching variable has U-shaped distributions, but the difference seems to be negligible when the matching variable follows a *Beta*(3.0, 3.0) distribution. This result is consistent with Levin and Paik [3] where estimation bias was decreased at a unimodal distribution, but maximized at a U-shaped distribution. Similar results were observed for the 1:4 matched design (data not shown). Since the 1:4 matched design basically increased the sample size of controls, overall selection performance was improved for both plogit and pclogit models but the difference between them was slightly reduced, compared with the 1:1 matched pairs. This simulation results indicate the superiority of conditional logistic analysis over unconditional logistic analysis in high-dimensional variable selection problems for matched case-control studies. In conclusion, the ordinary logistic fit ignoring matching for retrospective matched designs should be avoided for variable selection as well as estimation.

We conducted additional simulations for high-dimensional data with 20,000 CpG sites and 100 paired samples, where all simulation settings remain the same as the previous simulation except for $p = 20,000$ and $K = 100$. Since we increased $p = 20,000$ while kept the number of causal sites fixed at 40, we have extremely sparse association signals (only 0.2% variables are associated with the outcome). Therefore, the comparison of TNR is not meaningful because it is always greater than 99.49% ($1 - 100/19,960 = 0.9949$) for up to 100 any selected variables. We compared TPR and MCC of both plogit and pclogit instead with different number of top ranked variables. Figure 1 displays the averaged TPR and MCC over 50 simulation replicates when 1 to 100 top ranked variables are selected. The results of

plogit and pclogit are very similar to those from the previous simulation with fewer CpG sites, where pclgoit overall outperformed plogit except when the matching variable formed a unimodal distribution.

3.2. Scenario II: graph-constrained variables

In the second type of simulation studies when predictors are graphically constrained, we investigated the selection performance of several different versions of our network-based regularization method for conditional logistic analysis. We showed that incorporating genetic network information into our penalized conditional logistic model is very helpful to identify causal genes when genes are truly correlated with each other according to the genetic pathway information. We did not compare to the unconditional logistic regressions as the first type of simulation studies have suggested advantages of conditional logistic regression in variable selection for matched data.

The number of genes was fixed at $p = 1,000$, but two different sample sizes were considered, i, e., the number of strata were set to $K = 50$ and 100 for the 1:1 matched pairs design. We created a network graph in Figure 2 that mimics a genetic regulatory pathway, which usually contains a few hub genes plus many other genes with a few links. In the simulation we assumed that each genetic network consists of 100 genes corresponding to the pre-specified network in Figure 2. Therefore, we have 10 disjointed genetic networks including 1,000 genes. We further assumed one genetic network out of the 10 networks is disease related, where only 45 genes out of the 100 genes in this network are causal genes. Figure 2 depicts these 45 colored genes that consist of 4 groups of linked genes plus the centered gene in the network. We denoted these 4 groups of causal genes as g_1, g_2, g_3, and g_4, respectively.

Next, we assigned the strength of true signals so that a gene with more genetic links can have stronger association signals than a gene with less links. The mean vector of cases $\mu = (\mu_1,\ldots, \mu_p)^T$ was defined as

$$\mu_u = \begin{cases} \delta\sqrt{d_u}/3, & \text{if } u \in g_1 \text{ or } g_2 \\ -\delta\sqrt{d_u}/3, & \text{if } u \in g_3 \text{ or } g_4 \\ \delta, & \text{if centered gene} \end{cases}$$

for $1 \le u \le 45$, and $\mu = 0$ for $u > 45$. In this setting, the genes in the first 2 causal gene groups are positively associated with a disease while the genes in the other causal gene groups are negatively associated with a disease. Therefore, all linked genes have the same signs in their regression coefficients except for the centered gene. But, this is an oversimplified setting compared to real data situations, so we randomly picked two genes not linked with each other from groups g_1 through g_4, and reversed their signs.

We then generated a covariance matrix based on the specified network graph using a Gaussian graphical model [24], where nonzero entries of inverse covariance matrix correspond to links between two genes of a network graph. We refer the reader to Peng et al. [25] for a detailed procedure to generate a covariance matrix given a network graph. After

we obtained a mean vector of μ and a covariance matrix of Σ, we simulated the multivariate normal data X using (4), where we used the same parameter setting of δ and the distribution of Z as in scenario I.

We considered three different versions of our network-based regularization procedures. For the first model, we replaced both S and L in (2) by an identity matrix, and then it was simply reduced to the elastic-net (ElasticNet) procedure [26]. The second version employed a fully connected network model to clustered genes within the same network group and was defined as the group network model (GroupNet). While ElasticNet assumes no links among genes, GroupNet assumes that each of the 100 genes within a genetic network forms one group where all genes are linked with each other. Note that these two procedures did not fully utilize the true genetic network structure in Figure 2. The third model is the graph-constrained (GraphNet) procedure where the Laplacian matrix L in (2) was computed based on the true network graph with S being preliminarily estimated by ridge regression.

The simulation results are displayed in Table 2. All selection criteria except AUC were calculated based on the top 50 selected genes ranked by selection probabilities using 150 resamplings. Also, simulation was repeated 100 times to summarize their averages and standard errors in this study. It appears that ElasticNet and GroupNet show relatively poor performance on selecting causal genes, compared with GraphNet in both 50 and 100 paired samples. Particularly, GraphNet has around 11~17% higher true positive rates than the others in the 50 paired samples. This means that GraphNet can detect around 5~7 more causal genes than the other methods on average when top 50 genes are repeatedly selected through simulation replications. When the sample size was doubled, this difference was slightly decreased, but the true positive rates of GraphNet are still approximately 8~11% higher.

Here, we demonstrated that the network-based procedure incorporating a prior network information overwhelms the procedure ignoring the information when data are actually correlated according to a pre-specified network in matched case-control designs. Therefore, we expect that utilizing genetic pathways for our penalized conditional logistic model would identify more causal genes. However, we note that, in reality it is hard to completely capture true correlation patterns of high-dimensional genomic data since biological pathway information is usually limited to partial genes.

4. Data Analysis

We applied our penalized conditional logistic regression model to a genome-wide DNA methylation study on HCC where DNA methylation levels of tumor and adjacent non-tumor tissues from HCC patients were investigated using the Illumina Infinium HumanMethylation27 Beadchip [1]. The Illumina Infinium HumanMethylation27 BeadChip interrogates 27,578 highly informative CpG sites covering 14,495 genes. The methylation β-values were generated using the Illumina BeadStudio software. After removal of the sex chromosomes and samples with low CpG coverage, 26,486 autosomal CpG sites were left for 62 matched tumor and non-tumor pairs. The data is publicly available at the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/).

We further removed CpG sites with any missing β-values since the regularization procedures cannot be performed on predictors with missing values. Moreover, CpG sites that have missing gene annotation were also removed. We ended up with 22,884 CpG sites from 12,842 genes, where 3,793 and 8,844 genes have 1 and 2 CpG sites, respectively. Only 174 genes have 3 to 9 CpG sites, and the rest 31 genes have 10 to 28 CpG sites.

We analyzed this matched DNA methylation data in two different ways; either utilize biological group information such as correlated DNA methylation levels of multiple CpG sites within a gene or utilize graph information such as graphically constrained gene-level DNA methylation within a genetic pathway. In the first analysis, we focused on identifying CpG sites that are differentially methylated between tumor and adjacent non-tumor tissues of HCC patients where we employed the group network regularization method that assumes multiple CpG sites within one gene are all linked with each other [2] similarly as in simulation scenario I. In the second analysis, we focused on identifying genes that are differentially methylated between tumor and adjacent non-tumor tissues of HCC patients where we employed the graph-constrained (GraphNet) procedure that utilizes genetic pathway information and is shown to outperform other methods in simulation scenario II. Selection probability of each CpG site in the first analysis and selection probability of each gene in the second analysis were computed based on 150 resamplings. In the second analysis, we employed the same genetic pathway data as was used in Chen et al. [5]'s GWAS of Crohn's disease, where 3,735 genes over 350 pathways were combined from BioCarta (http://www.biocarta.com) and KEGG [27]. To use this pathway information, we screened and only kept those genes that can be mapped to the pathways. We ended up with a network consisting of 2,801 genes and 11,988 links. As there are 1 to 28 CpG sites per gene, we extracted the first principal component (PC) of the multiple CpG sites within a gene on logit transformed (to ensure we have the same gene-level methylation scale especially for genes with a single CpG site) β-values to achieve a gene-level overall DNA methylation measure. Finally, we used the first PC for each gene to conduct gene selection within genetic pathways.

Table 3 displays the top 25 CpG sites ranked by selection probabilities among 22,884 CpG sites along with the names of genes they belong to under **Analysis I** and the top 25 genes ranked by selection probabilities among 2,801 genes under

Analysis II. The unadjusted *p*-values from the paired t-test as used by Shen et al. [1] and their ranking among the 22,884 CpG sites were also listed under **Analysis I**. Note that the top 25 selected CpG sites in Analysis I not only are all significant at a Bonferroni adjusted level of significance ($p < 0.05/22,884$) using the paired t-test but also are ranked high among all significant sites. Compared to Table 2 in Shen et al. [1] which presented top 20 hyper- and top 20 hypo-methylated CpG sites and their corresponding genes in HCC tumor tissues compared to adjacent non-tumor tissues based on adjusted *p*-values, 15 are also in our top 25 selected CpG sites based on the selection probabilities. Among the 10 CpG sites that are in our top 25 list but not in Table 2 in Shen et al. [1], 3 sites are from genes that appeared in Table 2 in Shen et al. [1]. Among the rest 7 CpG sites that are in our top 25 list but not in Table 2 in Shen et al. [1], three are from genes SEMA3B and WFDC1 that are suggested to be tumor suppressor genes (http://www.genecards.org).

To examine the top 25 genes ranked by selection probabilities listed under **Analysis II** in Table 3, we note that only partial genes that are available in the pathway network (2,801 genes) were utilized in Analysis II, making the result not completely comparable with that in the original paper by Shen et al. [1] which was based on 14,495 genes. Four genes (BMP4, CDKN2A, CFTR, and RASSF1) from the gene list of the top 20 hyper-methylated genes and another 4 genes (AKT3, PAX4, CD1B, and CYP11B1) from the gene list of the top 20 hypo-methylated genes in Table 2 in Shen et al. [1] were in the pathway network, thus were included in Analysis II. Five genes (CDKN2A, CFTR, CYP11B1, CD1B, and PAX4) out of the 8 genes were selected by Analysis II in the top 25 gene list. In our top 25 gene list in Analysis II, we also selected several new genes that are known to be very important in HCC. For example, gene HOXA9 was suggested to be methylated more frequently in HCC cases that are HBV-positive [28]. Gene CASP8 was identified as significantly hyper-methylated in liver tissues from HCC patients, compared with normal liver tissues [29]. Also, gene VIM has been recently identified to be associated with HCC using microRNA expression data [30]. In another paper [31], gene SPPR3 showed different methylation levels between nontumorous cirrhotic tissue and normal liver tissue in HCC patients. Finally, the human leukocyte antigen-G (HLA-G) gene is also suggested to be associated with HCC prognosis in a Chinese population based on indel polymorphism data [32].

We also noticed some limitations. Our Analysis II never selected gene RASSF1, which is known to be a tumor suppressor gene and was also in the top 20 hyper-methylated gene in the original paper by Shen et al. [1]. To further investigate, we found that the first PC of RASSF1 gene with 8 CpG sites only explains 44% of the total variation. The second PC explains 34% variation, which suggests the two PCs may be combined to represent gene RASSF1 better. However, our analysis incorporating genetic pathway data cannot include two PCs per gene which is also a common limitation in using PC in genetic association studies.

5. Discussion

In spite of the common usage of the matched designs in epigenetic studies, there are few variable selection methods specifically designed for matched case-control studies, compared with those for unmatched case-control studies. In this article, we introduced a penalized conditional logistic model with a network-based penalty for the analysis of matched high-dimensional data with focus on DNA methylation data. But the method is readily applied to matched studies with genomic data such as gene expression data constrained by a genetic pathway. The proposed method can also be readily applied to the 1:1 matching strategy described by Yu and Deng [33] to analyze case-parent trio data or sibship data with slight modification, where high-dimensional SNP data might be grouped by genes to encourage a gene effect.

In simulation studies we have demonstrated that the proposed conditional logistic model much improved selection performance compared to the unconditional logistic model ignoring matching for analysis of matched case-control data to select disease related sites. In the real data application of our penalized conditional logistic model to a genome-wide DNA methylation study with tumor and adjacent non-tumor tissues of HCC patients, we have not

only confirmed many original selected sites/genes, but also selected several new sites/genes that are known to be important in HCC or cancer.

In our Analysis II, we relied on gene-level principal components which is known to be limited. A hierarchical selection model that takes the three layers into account, namely, multiple CpG sites with in a gene, multiple genes within a pathway, to select disease related CpG sites is our current research topic. Our network-based regularization procedure for the conditional logistic model was implemented into an R package pclogit which can be downloaded at http://www.columbia.edu/~sw2206.

Acknowledgments

This work was supported by National Institute of Health (R03 CA150140-01, R01 ES005116-19A1). The authors want to thank the Yale High Performance Computing center for providing computing facility. We would also like to thank our collaborators Drs. Regina Santella, Jing Shen, Yu-Jing Zhang, Maya Kappil, Hui-Chen Wu, Qian Wang from Department of Environmental Health Sciences, Columbia University, Drs. Muhammad G. Kibriya, Farzana Jasmine, Habib Ahsan from Department of Health Studies, University of Chicago, and Drs. Po-Huang Lee, Ming-Whei Yu, Chien-Jen Chen from National Taiwan University. Finally, we want to thank Dr. Min Chen for providing genetic pathway data and Dr. Mengling Liu for help discussions.

References

1. Shen J, Wang S, Zhang YJ, Kappil M, Wu HC, Kibrya MG, Wang Q, Jasmine F, Ahsan H, Lee PH, Yu MW, Chen CJ, Santella RM. Genome-wide DNA methylation profiles in hepatocellular carcinoma. Hepatology. 2012; 55(6):1799–1808. [PubMed: 22234943]
2. Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data analysis with case-control studies. Bioinformatics. 2012; 28(10):1368–1375. [PubMed: 22467913]
3. Levin B, Paik MC. The unreasonalbe effectiveness of a biased logistic regression procedure in the anaylsis of pair-matched case-contorl studies. Journal of Statistical Planning and Inference. 2001; 96(1):371–385.
4. Hansson L, Khamis HJ. Matched samples logistic regression in case-control studies with missing values: when to break the matches. Statistical Methods in Medical Research. 2008; 17(6):595–607. [PubMed: 18375456]
5. Chen M, Cho J, Zhao H. Incorporating biological pathways via a markov random field model in genome-wide association studies. PLoS Genetics. 2011; 7(4):e1001353. [PubMed: 21490723]
6. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24(9):1175–1182. [PubMed: 18310618]
7. Li C, Li H. Variable selection and regression analysis for covariates with a graphical structure with an application to genomics. The Annals of Applied Statistics. 2010; 4(3):1498–1516. [PubMed: 22916087]
8. Pan W, Xie B, Shen X. Incorporating predictor network in penalized regression with application to microarray data. Biometrics. 2010; 66(2):474–484. [PubMed: 19645699]
9. Fleiss, JL.; Levin, B.; Paik, MC. Statistical Methods for Rates and Proportions. Wiley; 2003.
10. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics. 2005; 21(13):3001–3008. [PubMed: 15814556]
11. Engler D, Li Y. Survival analysis with high-dimensional covariates: An application in microarray studies. Statistical Applications in Genetics and Molecular Biology. 2009; 8(1)
12. Wang S, Nan B, Zhou N, Zhu J. Hierarchically penalized Cox regression with grouped variables. Biometrika. 2009; 96 (2):307–322.
13. Goeman J. L1 penalized estimation in the cox proportional hazards model. Biometrical Journal. 2010; 52(1):70–84. [PubMed: 19937997]

14. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of Statistical Software. 2011; 39(5):1–13. [PubMed: 21572908]
15. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society B. 2006; 68(1):49–67.
16. Meier L, Geer S, Bühlmann P. The group lasso for logistic regression. Journals of the Royal Statistical Society B. 2008; 70(1):53–71.
17. Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. The Annals of Applied Statistics. 2007; 1(2):302–332.
18. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. The Annals of Applied Statistics. 2008; 2(1):224–244.
19. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society B. 2010; 72(4):417–473.
20. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000; 16(5):412–424. [PubMed: 10871264]
21. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press; 2003.
22. Kuan P, Wang S, Zhou X, Chu H. A statistical framework for illumina DNA methylation arrays. Bioinformatics. 2010; 26 (22):2849–2855. [PubMed: 20880956]
23. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. Biostatistics. 2012; 13(1):166–178. [PubMed: 21685414]
24. Whittaker, J. Graphical Models in Applied Mathematical Multivariate Statistics. Wiley; 1990.
25. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association. 2009:104.
26. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B. 2005; 67(2):301–320.
27. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000; 28(1):27–30. [PubMed: 10592173]
28. Feng Q, Stern J, Hawes S, Lu H, Jiang M, Kiviat N. DNA methylationchanges in normallivertissues and hepatocellularcarcinoma with differentviralinfection. Experimental and Molecular Pathology. 2010; 88(2):287–292. [PubMed: 20079733]
29. Nishida N, Nagasaka T, Nishimura T, Ikai I, Boland C, Goel A. Aberrant methylation of multiple tumor suppressor genes in aging liver, chronic hepatitis, and hepatocellular carcinoma. Hepatology. 2008; 47(3):908–918. [PubMed: 18161048]
30. Furuta M, Kozaki K, Tanaka S, Arii S, Imoto I, Inazawa J. miR-124 and miR-203 are epigenetically silenced tumor-suppressive microRNAs in hepatocellular carcinoma. Carcinogenesis. 2010; 31(5):766–776. [PubMed: 19843643]
31. Ammerpohl O, Pratschke J, Schafmayer C, Haake A, Faber W, Kampen O, Brosch M, Sipos B, Schonfels W, Balschun K, Rocken C, Arlt A, Schniewind B, Kalthoff JGH, Neuhaus P, Stickel F, Schreiber S, Becker T, Siebert R, Hampe J. Distinct DNA methylation patterns in cirrhotic liver and hepatocellular carcinoma. International Journal of Cancer. 2012; 130(6):1319–1328.
32. Jiang Y, Chen S, Jia S, Zhu Z, Gao X, Dong D, Gao Y. Association of HLA-G 3′ UTR 14-bp insertion/deletion polymorphism with hepatocellular carcinoma susceptibility in a chinese population. DNA and Cell Biology. 2011; 30 (12):1027–1032. [PubMed: 21612396]
33. Yu Z, Deng L. Pseudosibship methods in the case-parents design. Statistics in Medicine. 2011; 30(27):3236–3251. [PubMed: 21953439]

Appendix

In this Appendix, we will show how to compute selection probabilities in the penalized conditional logistic regression model.

1. Set a and $\lambda \in \Lambda$, where Λ is a regularization parameter space.
2. Randomly sample a subset of strata $\{1,\ldots, K\}$ that has a size of $\langle K/2 \rangle$ without replacement, where $\langle x \rangle$ is the largest integer no greater than x.
3. Apply the penalized conditional logistic model to the subset of strata, given a and λ. Denote $\hat{\beta}_u^l$ by the u-th regression coefficient estimate based on the l-th subset of strata.
4. Repeat (2)-(3) L times.
5. The selection probability of the u-th predictor can be defined as

$$\hat{p}_u = \max_{\lambda \in \Lambda} \frac{1}{L} \#\{l \leq L : \hat{\beta}_u^l) \neq 0\}.$$

Selection results rarely depend on the choice of Λ, including a singleton, according to Meinshausen and Buhlmann [19]. Similar to Simon et al. [14], we set a sequence of λ such that regression coefficients have some non-zero values each λ. We fixed $a = 0.1$ in this manuscript since a small a value is generally known as efficient for highly correlated data [2].

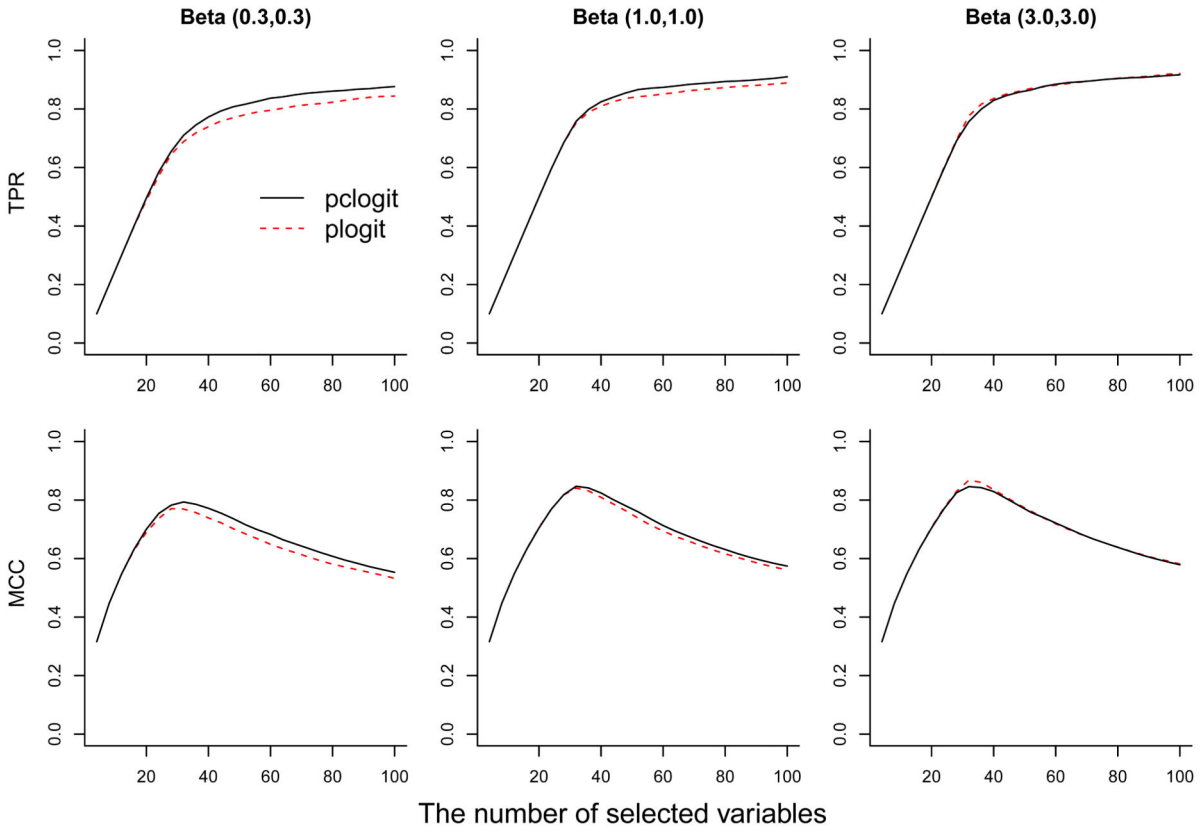

Figure 1.

The penalized ordinary logistic (plogit) and penalized conditional logistic (pclogit) models were compared along with different number of selected variables that are ranked by selection probabilities, while allowing a matching variable to follow a *Beta* distribution with different parameter settings. Averaged true positive rate (TPR) in the upper panels and Mathews Correlation coefficient (MCC) in the lower panels are displayed.

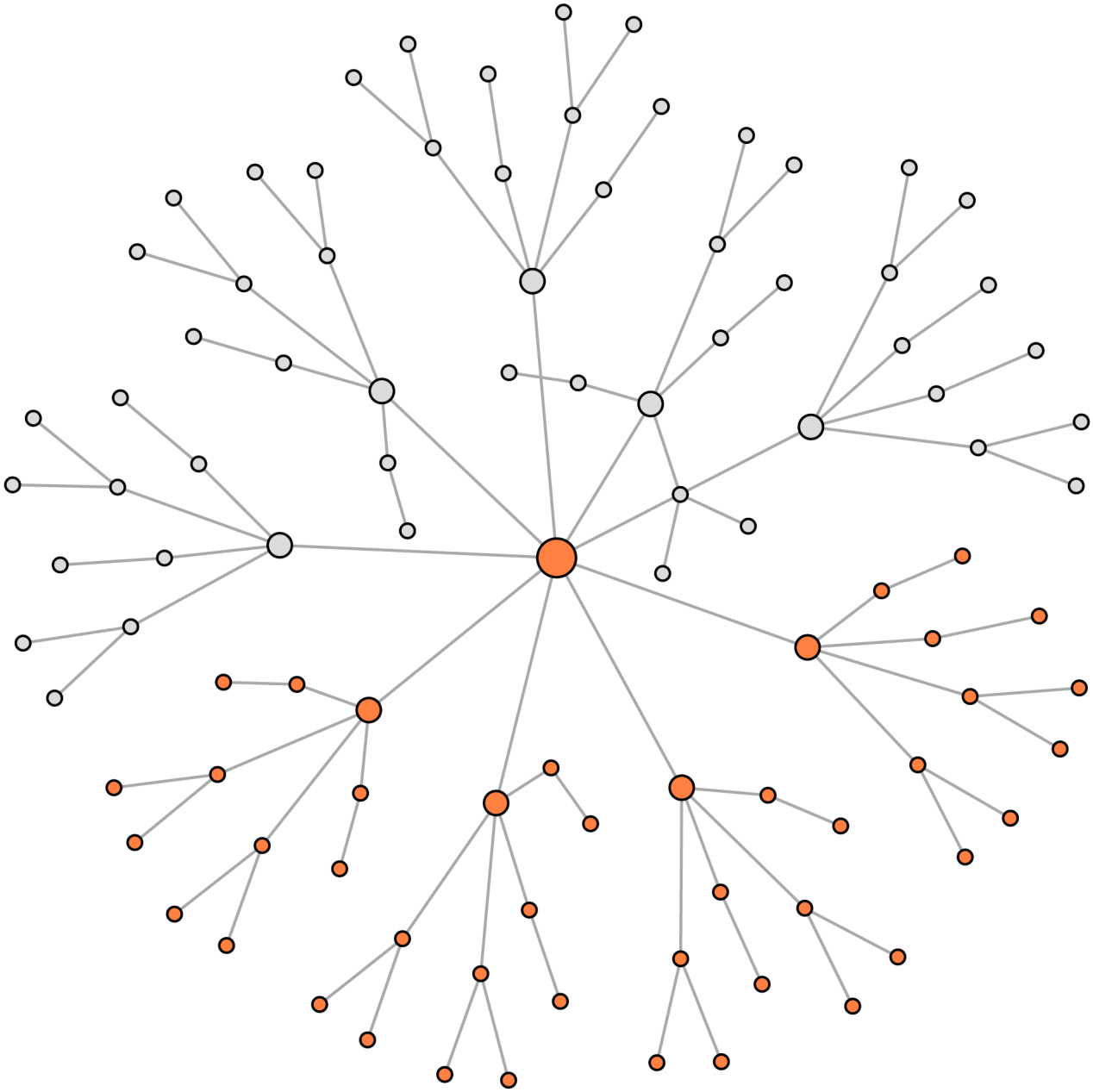

Figure 2.
An example of a network graph with 100 genes used for scenario II in simulation. The colored 45 genes are assumed to be causal genes and they consist of one centered gene plus 4 different groups of genes, where each group has 11 genes.

Table 1

Scenario I; The penalized ordinary logistic (plogit) and penalized conditional logistic (pclogit) models were compared based on top 40, 60, and 80 variables (T40, T60, and T80) ranked by selection probabilities, allowing a matching variable to follow a *Beta* distribution with different parameter settings. Averaged true positive rate (TPR), true negative rate (TNR), Mathews Correlation Coefficient (MCC), and area under the ROC curve (AUC) are present with their standard errors.

	Method	*Beta*(0.3, 0.3)	*Beta*(0.5, 0.5)	*Beta*(1.0, 1.0)	*Beta*(3.0, 3.0)
TPR	plogit.T40	0.6700 (0.0075)	0.6842 (0.0062)	0.7578 (0.0089)	0.7917 (0.0051)
	plogit.T60	0.7655 (0.0086)	0.7785 (0.0065)	0.8565 (0.0077)	0.8660 (0.0053)
	plogit.T80	0.8128 (0.0084)	0.8232 (0.0059)	0.8825 (0.0076)	0.8905 (0.0051)
	pclogit.T40	0.7678 (0.0077)	0.7725 (0.0051)	0.7760 (0.0096)	0.7907 (0.0057)
	pclogit.T60	0.8492 (0.0075)	0.8510 (0.0047)	0.8702 (0.0113)	0.8738 (0.0057)
	pclogit.T80	0.8852 (0.0068)	0.8945 (0.0035)	0.9075 (0.0076)	0.9100 (0.0051)
TNR	plogit.T40	0.9862 (0.0003)	0.9868 (0.0003)	0.9899 (0.0004)	0.9913 (0.0002)
	plogit.T60	0.9694 (0.0004)	0.9699 (0.0003)	0.9732 (0.0003)	0.9736 (0.0002)
	plogit.T80	0.9505 (0.0003)	0.9510 (0.0002)	0.9534 (0.0003)	0.9538 (0.0002)
	pclogit.T40	0.9903 (0.0003)	0.9905 (0.0002)	0.9907 (0.0004)	0.9913 (0.0002)
	pclogit.T60	0.9729 (0.0003)	0.9730 (0.0002)	0.9738 (0.0005)	0.9739 (0.0002)
	pclogit.T80	0.9536 (0.0003)	0.9539 (0.0001)	0.9545 (0.0003)	0.9546 (0.0002)
MCC	plogit.T40	0.6562 (0.0078)	0.6711 (0.0065)	0.7477 (0.0093)	0.7831 (0.0053)
	plogit.T60	0.6064 (0.0074)	0.6176 (0.0056)	0.6846 (0.0066)	0.6928 (0.0045)
	plogit.T80	0.5513 (0.0063)	0.5592 (0.0045)	0.6038 (0.0057)	0.6098 (0.0039)
	pclogit.T40	0.7581 (0.0081)	0.7630 (0.0053)	0.7667 (0.0100)	0.7820 (0.0059)
	pclogit.T60	0.6784 (0.0064)	0.6799 (0.0041)	0.6964 (0.0097)	0.6994 (0.0049)
	pclogit.T80	0.6059 (0.0051)	0.6128 (0.0027)	0.6226 (0.0057)	0.6245 (0.0038)
AUC	plogit	0.9507 (0.0035)	0.9501 (0.0023)	0.9711 (0.0023)	0.9764 (0.0016)
	pclogit	0.9719 (0.0025)	0.9693 (0.0021)	0.9772 (0.0025)	0.9750 (0.0016)

Stat Med. Author manuscript; available in PMC 2014 May 30.

Table 2

Senario II; The elastic-net (ElasticNet), group-network (GroupNet), and graph-constrained (GraphNet) regularization procedures are compared based on top 50 variables ranked by selection probabilities, allowing a matching variable to follow a *Beta* distribution with different parameter settings. Averaged true positive rate (TPR), true negative rate (TNR), Mathews Correlation Coefficient (MCC), and area under the ROC curve (AUC) are present with their standard errors.

		50 paired samples			100 paired samples		
	Method	***B*(0.3, 0.3)**	***B*(1.0, 1.0)**	***B*(3.0, 3.0)**	***B*(0.3, 0.3)**	***B*(1.0, 1.0)**	***B*(3.0, 3.0)**
TPR	ElasticNet	0.693 (0.005)	0.728 (0.006)	0.728 (0.005)	0.878 (0.004)	0.902 (0.004)	0.903 (0.004)
	GroupNet	0.690 (0.004)	0.722 (0.005)	0.727 (0.006)	0.874 (0.003)	0.904 (0.004)	0.889 (0.003)
	GraphNet	0.850 (0.004)	0.881 (0.004)	0.893 (0.005)	0.970 (0.003)	0.985 (0.002)	0.993 (0.001)
TNR	ElasticNet	0.980 (0.000)	0.982 (0.000)	0.982 (0.000)	0.989 (0.000)	0.990 (0.000)	0.990 (0.000)
	GroupNet	0.980 (0.000)	0.982 (0.000)	0.982 (0.000)	0.989 (0.000)	0.990 (0.000)	0.990 (0.000)
	GraphNet	0.988 (0.000)	0.989 (0.000)	0.990 (0.000)	0.993 (0.000)	0.994 (0.000)	0.994 (0.000)
MCC	ElasticNet	0.640 (0.005)	0.675 (0.006)	0.675 (0.005)	0.825 (0.004)	0.849 (0.004)	0.850 (0.004)
	GroupNet	0.638 (0.004)	0.669 (0.005)	0.675 (0.006)	0.821 (0.003)	0.851 (0.004)	0.835 (0.003)
	GraphNet	0.797 (0.004)	0.827 (0.004)	0.839 (0.005)	0.917 (0.003)	0.932 (0.002)	0.939 (0.001)
AUC	ElasticNet	0.941 (0.002)	0.949 (0.002)	0.961 (0.001)	0.986 (0.001)	0.993 (0.001)	0.993 (0.000)
	GroupNet	0.942 (0.002)	0.949 (0.002)	0.960 (0.001)	0.984 (0.001)	0.992 (0.001)	0.994 (0.000)
	GraphNet	0.984 (0.001)	0.989 (0.001)	0.992 (0.001)	0.998 (0.000)	1.000 (0.000)	1.000 (0.000)

Stat Med. Author manuscript; available in PMC 2014 May 30.

Table 3

Analysis I; the top 25 CpG sites selected by their selection probabilities (Sel.prob) using the group network regularization procedure. The corresponding gene names are present along with their unadjusted *p*-values based on the paired t-test and rankings (p-value rank) among 22,884 CpG sites using paired t-test. **Analysis II**. the top 25 genes selected by their selection probabilities (Sel.prob) using the graph-constrained regularization procedure.

	Analysis I					Analysis II	
Rank	IlmnID	GENE	Sel.prob	p-value	p-value rank	GENE	Sel.prob
1	cg14310034	BMP4	1.0000	2.022732E-20	2	CDKN2A	1.0000
2	cg21643045	CCL20	1.0000	6.417129E-19	4	CFTR	1.0000
3	cg05684891	DAB2IP	1.0000	1.353434E-21	1	CYP11B1	1.0000
4	cg15014458	LYPD3	1.0000	2.207219E-13	89	HOXA9	1.0000
5	cg25340403	LYPD3	1.0000	9.074690E-15	42	IRAK3	1.0000
6	cg24816455	SEMA3B	1.0000	6.819552E-14	73	MAP3K14	1.0000
7	cg04786857	SPDY1	1.0000	6.771017E-19	5	CASP8	0.9933
8	cg12680609	ZFP41	1.0000	1.227676E-19	3	CD1B	0.9933
9	cg24432073	CDKL2	0.9933	3.033569E-17	8	VIM	0.9933
10	cg09099744	CDKN2A	0.9933	3.279436E-18	6	CR1	0.9867
11	cg14988503	CDKL2	0.9867	4.962911E-15	35	LAMB3	0.9867
12	cg10143146	COL11A2	0.9800	2.237076E-14	55	PAX4	0.9800
13	cg15868302	FOXD2	0.9667	1.088636E-15	24	SPRR3	0.9800
14	cg23865698	WFDC1	0.9467	1.086906E-14	45	INS	0.9600
15	cg03975694	ZNF540	0.9400	7.650334E-17	10	AKR1B1	0.9533
16	cg14911395	SEMA3B	0.9333	1.902788E-10	376	RALGDS	0.9533
17	cg09120035	CYP11B1	0.9267	9.779697E-16	22	PRKCB1	0.9400
18	cg25934198	FLJ36268	0.9067	3.891966E-11	272	HLA-G	0.9333
19	cg03292149	CBLC	0.8867	2.975328E-11	256	DOCK2	0.9200
20	cg11377136	PKDREJ	0.8867	1.955923E-16	14	GK2	0.9200
21	cg01076838	TMEM86B	0.8667	1.112359E-14	46	CAPN2	0.9133
22	cg13615963	CCR6	0.8467	2.717370E-15	31	NPY	0.9133
23	cg03602500	FLJ00060	0.8467	5.154408E-16	18	INA	0.9000
24	cg25564800	KPNA1	0.8467	1.434292E-15	25	PARVG	0.9000
25	cg10895543	CDKN2A	0.8400	8.921449E-16	20	EIF4B	0.8933