BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Network-based classification of recurrent endometrial cancers using high-throughput DNA methylation data

Jianhua Ruan [1,*,†], Md. Jamiul Jahid [1,*], Fei Gu [2], Chengwei Lei [1], Yi-Wen Huang [3], Ya-Ting Hsu [2], Paul J. Goodfellow [4], Chun-Liang Chen [2], Tim H.-M. Huang [2,†]

[1]Department of Computer Science, University of Texas, San Antonio, TX 78249, USA, [2]Department of Molecular Medicine, University of Texas Health Science Center, San Antonio, TX 78245, USA, [3]Department of Obstetrics and Gynecology, Medical College of Wisconsin, Milwaukee, WI 53226, USA, [4]Department of Surgery, Washington University School of Medicine, St. Louis, MO 63110, USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: DNA methylation plays important roles in cancer, which is a complex disease involving many genes. However, by far DNA methylation analysis has not been integrated with the gene / protein networks that regulate various biological processes within the cell. Here, we developed a novel computational method to analyze whole-genome DNA methylation data for endometrial tumors within the context of a human protein-protein interaction (PPI) network, in order to identify subnetworks as potential epigenetic biomarkers for predicting tumor recurrence. Our method consists of the following steps. First, differentially methylated (DM) genes between recurrent and non-recurrent tumors are identified and mapped onto a human PPI network. Then, a PPI subnetwork consisting of DM genes and genes that are topologically important for connecting the DMs on the PPI network, termed epigenetic connectors (ECs), are extracted using a Steiner-tree based algorithm. Finally, a random-walk based machine learning method is used to propagate the DNA methylation scores from the DMs to the ECs, which enables the ECs to be used as features in a support vector machine classifier for predicting recurrence.

Results: While the DMs are not enriched in any cancer-related pathways, the ECs are enriched in many well-known tumorgenensis and metastasis pathways and include known epigenetic regulators. Moreover, combining the DMs and ECs significantly improves the accuracy for classifying recurrence. Therefore, the network-based method is effective in identifying a subnetwork consisting of both differentially methylated genes and other important non-differentially methylated genes which are nevertheless important for the understanding and prediction of tumor recurrence.

Contact: jruan@cs.utsa.edu, huangt3@uthscsa.edu

1 INTRODUCTION

Increasing evidence shows that DNA methylation plays a significant role in cancer, from the silencing of tumor suppressors to the activation of oncogenes and the promotion of metastasis, as well as the development of drug resistance (Huang and Esteller, 2010; Kulis and Esteller, 2010). Though not altering DNA sequence itself, this chemical modification may change chromatin structure that renders the accessibility of promoters to transcriptional machinery and regulate gene expression (Huang and Esteller, 2010; Kulis and Esteller, 2010).

The advent of next-generation sequencing technology combined with effective DNA methylation capture techniques is providing an unprecedented opportunity for a system-level understanding of methylation changes occurring in cancer and holds the promise of establishing epigenetic biomarkers for accurate cancer diagnosis and prognosis (Robinson *et al*., 2010; Serre *et al*., 2010; Huang and Esteller, 2010). However, by far large-scale DNA methylation analysis has not been integrated with the gene / protein networks that regulate gene expression and signal transduction within the cell. In transcriptomics-based cancer studies, it is frequently observed that individuals of the same phenotype may share similar expression patterns on the pathway level rather than the individual gene level (Radivojac *et al*., 2008; Jonsson and Bates, 2006; Wang *et al*., 2011; Vidal *et al*., 2011; Barabasi *et al*., 2011; Hidalgo *et al*., 2009; Yildirim *et al*., 2007; Goh *et al*., 2007; Li *et al*., 2010); therefore it has been increasingly realized that the underlying network topology must be considered, in order to obtain more useful biological insight from the "gene list" resulted from differential expression analysis and to improve the performance of cancer outcome predictions. With these critical observations and the fact that the majority of human genes have not been assigned to definitive pathways, many methods have been proposed to identify PPI subnetworks that are significantly differentially expressed or dysregulated in certain phenotypes as candidate pathway markers. This strategy has enabled systematic discovery of novel pathways associated with multiple diseases (Liu *et al*., 2007; Hwang *et al*., 2008; Geistlinger *et al*., 2011; Chowdhury *et al*., 2011; Keller *et al*., 2009; Ulitsky *et al*.,

*Equal contributions.

†To whom correspondence should be addressed

2010; Kim *et al*., 2011; Hung *et al*., 2010). In addition, several studies have proposed to use subnetworks as features to classify phenotypes, where such subnetworks are usually treated as meta-genes whose expression levels are defined as the mean expression levels of its nodes (Chuang *et al*., 2007; Chowdhury *et al*., 2011; Gatza *et al*., 2010; Lee *et al*., 2008).

Despite various levels of success, the current network-based approaches face a few challenges. First, it is known that some subnetworks are highly enriched in many disease subtypes, and therefore may represent downstream effects of the phenotypes (i.e., they are passengers rather than drivers). Second, some phenotypes are often associated with a huge number of genes that may have been regulated by a small number of master regulators, while the master regulators themselves may be buried in the long gene list and be overlooked, or are even not detected by the chosen profiling technique. Third, in many studies, subnetworks are used as meta-genes which may contain genes whose activities are not expected to change between phenotypes, and thus using their activities for classification can add noise and potentially deteriorate classification performance. Finally, subnetwork selection is usually done in a supervised manner prior to the training of classifiers, and therefore requires additional training samples and cannot be easily applied to clustering.

We believe that network-based analysis can also be beneficial or even necessary for epigenetic studies, as it is also true that epigenetic regulation of the expression levels of different genes on the same pathway may lead to the same phenotype; therefore, DNA methylation patterns should also be compared on pathway level rather than individual loci level. However, the above limitations to the existing subnetwork-based methods for transcriptomic studies also exist in epigenetic studies. Furthermore, as DNA methylation data have their own characteristics, it is not known whether the existing methods can be easily applied to identify epigenetic subnetwork markers.

In this study, we propose a novel computational method to analyze whole-genome DNA methylation data within the context of a human protein-protein interaction (PPI) network, and to identify subnetworks as potential biomarkers for predicting tumor recurrence. Our method consists of the following steps. First, differentially methylated (DM) genes between recurrent and non-recurrent tumors are identified and mapped onto a human PPI network. Then, a PPI subnetwork consisting of DM genes and genes that are topologically important for connecting the DM genes on the PPI network, termed epigenetic connectors (ECs), are extracted using a graph algorithm for finding multiple Steiner trees (Jahid and Ruan, 2012). As the ECs themselves may not necessarily be differentially methylated, we propose a random-walk based machine learning method to propagate the DNA methylation scores from the DMs to the ECs, which effectively enables the ECs to be used as biomarkers. Finally, the DM genes and the ECs are combined to construct a support vector machine for classifying recurrent versus non-recurrent tumors.

Applying our method to our unpublished high-throughput DNA methylation data of a panel of 60 primary endometrial tumors and 12 normal controls determined by methyl-CpG binding domain-based capture coupled with massively parallel sequencing (MBDCap-seq), we identified a set of DM genes and ECs whose combination can predict three-year tumor recurrence with an accuracy at 82.9%, as compared to 73.4% using the DMs alone. Furthermore, the ECs are significantly enriched with KEGG pathways well-known to be involved in tumorgenensis and metastasis, and include several known epigenetic regulators, signifying the effectiveness of our approach.

2 MATERIALS AND METHODS

2.1 Raw data collection and processing

Endometrial tissue specimens (74 primary endometrioid endometrial tumors and 12 uninvolved controls) were obtained as part of our ongoing work on characterizing molecular alterations in endometrioid endometrial carcinomas and were described in a previous report (Huang *et al*., 2010). Global DNA methylation pattern of the tumors and controls were surveyed using mthyl-CpG binding domain-based capture (Rauch and Pfeifer, 2010) coupled with massively parallel sequencing (MBDCap-seq; Robinson *et al*. 2010; Serre *et al*. 2010). Briefly, methylated DNA was eluted by the MethylMiner Methylated DNA Enrichment Kit (Invitrogen) according to the manufacturers instructions. Eluted DNA was used to generate libraries for sequencing following the standard protocols from Illumina. MBDCap-seq libraries were sequenced using the Illumina Genome Analyzer II as per manufacturer's instructions. Image analysis and base calling were performed with the standard Illumina pipeline. Sequencing reads were mapped by ELAND algorithm. Unique reads were up to 36 base pair reads mapped to the human reference genome (hg18), with up to two mismatches. Reads in satellite regions were excluded due to the large number of amplifications. The methylation level was normalized based on the unique read numbers for each sample by a linear method.

The tumor differential methylation (TDM) score was then calculated for each of the 13,081 known promoter CpG islands for each tumor by comparing the average methylation level in a 8-kb window covering the CpG island in the tumor relative to normal controls using one-sample t-test. For a CpG island that is hypermethylated (over-methylated) in tumor relative to controls, the TDM score is calculated as $-\log 10(p)$, where p is the p-value resulted from the t-test. For a CpG island that is hypomethylated (under-methylated) in tumor relative to controls, the TDM score is calculated as $\log_{10}(p)$. In both cases, p-values greater than 0.01 are converted to 1 and as a result the corresponding TDM scores are zero.

Detailed analysis of the complete DNA methylation data will be published elsewhere, and the data will be submitted to Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo).

2.2 Epigenetic marker and epigenetic connector subnetwork selection

In this step of the analysis, patients that had persistent tumors, or had non-recurrent tumor but their last follow-ups were within three years after surgery were excluded. As a result, a total of 60 patients were available for analysis, among which 16 had recurrence within 3 years, and the remaining 44 are considered non-recurrent. This dataset contains 4214 CpG islands that has non-zero TDM scores for at least one patient.

To identify potential epigenetic markers for recurrence, genes whose promoter CpG islands are differentially methylated between the recurrent and non-recurrent patients were identified by

comparing the TDM scores for each CpG island in the recurrent vs non-recurrent tumors using two-sample t-test. Genes with a p-value < 0.02 are termed differentially methylated (DM) genes.

Next, we mapped the DM genes to the human protein-protein interaction network obtained from HPRD (Release 9) (Prasad *et al*., 2009). We only took the largest connected component of the network, which contains 9,205 unique genes (official gene symbols) and 36,720 interactions.

We then sought to find a small number of genes that can connect the DM genes into a singly connected component in the PPI network. Intuitively, we are looking for the most parsimonious solution, that is, a spanning tree that connects the DM genes with the fewest additional genes. In graph theory, this maps to the well-known Steiner tree problem. Formally, the Steiner tree for an edge-weighted graph G = (V, E, w) and a subset of vertices S⊆V is a minimum-weight connected tree T, with vertices U⊆V and edges D⊆E that spans all vertices in S. Here the vertices in S are known as terminal vertices and U-S as Steiner vertices. For an unweighted graph G, the problem then becomes to find the minimum number of vertices that can connect all the vertices in S through a tree in G. The Steiner tree problem is NP-hard (Vo, 1992). We implemented a polynomial-time 2-approximation shortest path heuristic algorithm of Steiner tree problem (Rayward-Smith, 1983).

To improve the chance of finding all optimal Steiner vertices and also to cover as many alternative paths as possible, we designed a simple randomized algorithm to obtain multiple Steiner trees from a given set of input nodes. To this end, we assign to each edge of the PPI network a random weight between 0.99 and 1, and run the Steiner tree algorithm. These random weights effectively break ties, so that if there are two paths with the same weight in the original network, one path will be chosen randomly. This procedure was repeated multiple times with different random weights, until the total number of unique Steiner vertices converges approximately. In this work, the rate of new coming Steiner vertices reduced significantly after 200-300 iterations. We pooled the Steiner vertices in the 300 Steiner trees to obtain a set of unique genes, which we termed epigenetic connectors (ECs), as they play important roles in forming connections among the differentially methylated epigenetic markers.

2.3 Using EC genes as biomarkers

The ECs, by the way they are selected, are not differentially methylated between recurrent and non-recurrent tumors and therefore one would argue that they may not have any value as biomarkers. In our opinion, the ECs can be used in two ways in assisting with the classification of recurrence. First, the ECs contain a large portion of genes within the local neighborhood of DMs. While the DMs are generally universally differentially methylated between recurrent and non-recurrent, some of its neighbor genes (i.e., genes in the same pathway) may be differentially methylated at a level not considered statistically significant because, for example, they may only have TDM scores for a few patients. In this case, the combination of multiple weakly differentially methylated genes in the same pathway may also contribute to the separation of recurrent and non-recurrent. Second, as the ECs contain many topologically important genes connecting the DMs, e.g., hub genes with large connectivities or bottleneck genes connecting genes in different subregions of the network, they may be functionally important for the integrity of the DM subnetwork as well. For example, they may contain genes that are directly regulating the epigenetic changes of the DM genes or are functionally regulated by the DM genes. In either case, those genes, although not differentially methylated *per se*, may be considered as "proxies" to DM genes and can be utilized in classification indirectly.

To utilize as biomarkers EC genes falling into the first category, i.e., weakly differentially methylated genes, is relatively simple as most classification algorithms can handle combinations of features in some way. To deal with ECs genes falling into the second category, i.e., topologically important non-differentially methylated genes, we propose a novel machine learning algorithm to derive a score for each EC, based on the topological property of the gene in the network and its relative position with regard to all other genes. The method is adopted from the random walk with restart (RWR) algorithm (Tong *et al*., 2006) popular in machine learning and works as follows.

First we construct a modified, directed PPI subnetwork encompassing DMs and ECs so that the DM nodes do not have any incoming edges. In order words, the DM genes can only "pump" their TDM scores into the subnetwork but do not receive any scores. The ECs on the other hand have edges in both directions so they can act as both a donor and a receiver of TDM scores.

Formally, let A be the adjacency matrix of an unweighted, directed graph, where $A(i,j) = 1$ if there is an edge from node i to node j and 0 otherwise, and P be the row normalized adjacency matrix (or the transition probability matrix) defined on the graph, where $p_{ij} = \frac{A_{ij}}{\sum_j A_{ij}}$ is the transition probability from node i to node j. Assume that a random walker starts from a node v. At any discrete time point $k+1$, the probability for the random walker to take the path from node i to node j is $f_{ij}^{k+1}(v) = F_i^k(v)p_{ij}$, where $F_j^k(v)$ is the probability for the random walker to be at node j at time point k. Evidently, without considering the probability to revisit the starting node (hence "restart"), the probability to be at any node j for the random walk started at v is $F_j^k(v) = \sum_i f_{ij}^k(v)$ for any k. Now considering that the random walker always has a probability c to revisit the starting node v, we have $F_v^k(v) = (1-c)\sum_i f_{iv}^k(v) + c$, and $F_j^k(v) = (1-c)\sum_i f_{ij}^k(v)$ for all other $j \neq v$. In our case c is set to 0.5. This procedure is guaranteed to converge, as shown previously (Tong *et al*., 2006). The stationary probability vector $F^{\inf}(v)$, or simply denoted as $F(v)$, is the influence of node v on any node in the network. Evidently if v has no incoming edges, then $F_v(v) = c$ and $F_v(j) = 0$ for any $j \neq v$. This procedure is repeated using each node as a starting node and the vector $F(v)$ is pre-computed for every v.

Now consider a particular tumor, t. Let $s_i(t)$ be the TDM score of the i-th gene on the DM-EC subnetwork for t. The random walk based (RWB) score of gene i for t is calculated as: $r_i(t) = \sum_v s_v(t)F_i(v)$. It can be seen that for nodes with no incoming edges, $r_i(t) = cs_i(t)$. In other words, the effect of this random walk procedure to the DM genes is simply multiplying their TDM scores by a constant factor c. Therefore at the end we simply multiplied all the RWB scores by $1/c$ so that the TDM scores and RWB scores for the DM genes are equivalent.

2.4 Classification and performance evaluation

Support vector machine (SVM) classifiers were built using the implementation in WEKA 3.6.6 (Witten and Frank, 1999). A total

of six classifiers were built, using as features (1) TDM scores of DM genes, (2) TDM scores of EC genes, (3) TDM scores of DM and EC genes, (4) TDM scores of DM genes and 474 randomly selected genes on the PPI, (5) RWB scores of EC genes, and (6) TDM scores of the DM genes and RWB scores of the EC genes.

Classification performance was estimated 100 times using 10-fold cross-validation where iteratively one-tenth of the data were used for testing and nine-tenth were for training. Classification accuracy is defined as the percent of patients classified correctly. As the dataset has much more non-recurrent patients than recurrent patients, we also calculated kappa statistic, κ, which measures the agreement between the class labels and the predictions made by the classifier, corrected by the amount of agreement that may be achieved by chance (Landis and Koch., 1977). Formally, let TP, TN, FP, and FN be the numbers of true positive, true negative, false positive and false negative predictions made by a binary classifier, respectively, and N = TP + TN + FP + FN. The kappa static κ of the classifier is defined as $\kappa = \frac{A-C}{1-C}$, where $A = \frac{TP+TN}{N}$ is the fraction of correctly predicted instances and C is the expected percentage of instances that a classifier can predict correctly by chance, defined as

$$C = \frac{TP+FP}{N} \times \frac{TP+FN}{N} + \frac{TN+FN}{N} \times \frac{TN+FP}{N}.$$

Conventionally a kappa value over 0.75 is considered as excellent, 0.40 to 0.75 as good, and below 0.40 as poor.

3 RESULTS AND DISCUSSION

3.1 EC-subnetwork is enriched in cancer-related pathways

We identified 135 DM genes ($p < 0.02$, see Methods) connected by 474 EC genes (Fig. 1). Interestingly, most of the DM genes are hyper-methylated in non-recurrent cancers, while a small number of DM genes are hypo-methylated in recurrent cancers (Fig 2a). Unlike the DM genes, the EC genes are not differentially methylated between recurrent and non-recurrent cancers. Furthermore, many of the ECs have no non-zero TDM scores or have scores in only a few patients (Fig 2b), suggesting that DNA methylation change is not a main regulatory mechanism for them. Among the 474 EC genes, only 115 have a TDM score for at least one patient, and a mere of 8 have TMD scores for at least half of the patients (Fig 2c). In contrast, 68 of the 135 DM genes have TDM scores for at least half of the patients.

These genes were then used for KEGG pathway enrichment analysis by the Fisher's exact test, using the size of the PPI network as background for the ECs and the number of genes with non-zero TDM scores for the DMs. Remarkably, while the DM genes are not significantly enriched with any known KEGG pathways, the EC genes are significantly enriched with many KEGG pathways that are well known to be involved in tumorgenesis and metastasis, such as GnRH signaling pathway ($p < 1E-14$), ErbB signaling pathway ($p < 1E-12$), gap junction ($p < 1E-12$), Wnt signaling pathway ($p < 1E-8$) and TGF-β ($p < 1E-6$), among others (Table 1). Interestingly, neurotrophin signaling pathway ($p < 1E-14$) and calcium signaling pathway ($p < 1E-13$) are also among the top-enriched pathways.

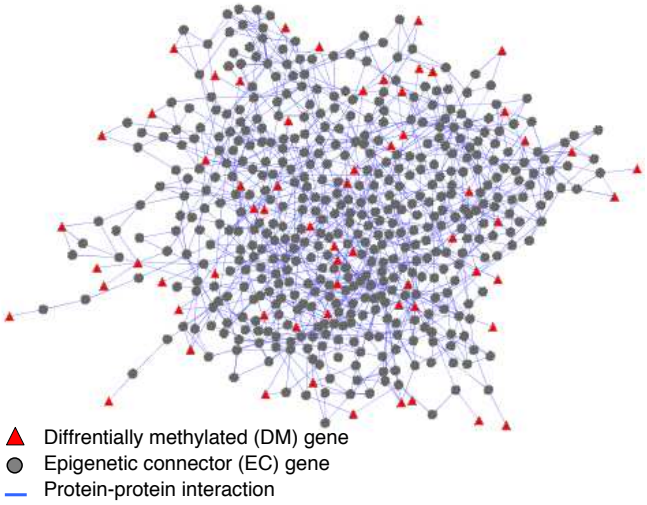


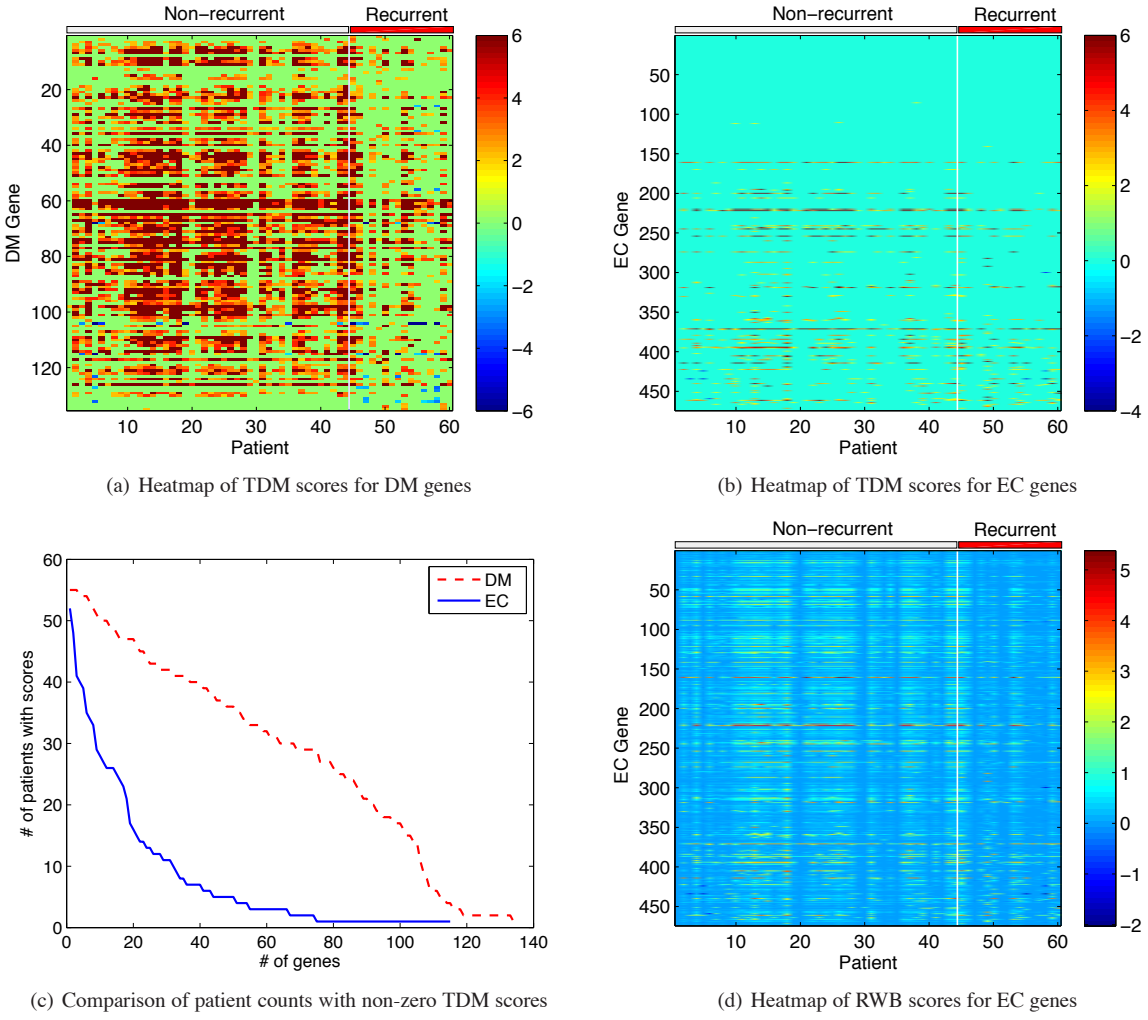
Fig. 1. A PPI subnetwork encompassing DM and EC genes.

Table 1. Enriched KEGG pathways in EC genes. Enrichment score is calculated as -log10(p-value).

KEGG Pathway	Score
hsa05200 Pathways in cancer	19.5
hsa04722 Neurotrophin signaling pathway	14.8
hsa04912 GnRH signaling pathway	14.0
hsa04020 Calcium signaling pathway	13.6
hsa04080 Neuroactive ligand-receptor interaction	13.1
hsa04012 ErbB signaling pathway	12.8
hsa04540 Gap junction	12.8
hsa04520 Adherens junction	12.4
hsa04062 Chemokine signaling pathway	11.3
hsa04510 Focal adhesion	10.5
hsa04010 MAPK signaling pathway	10.4
hsa04360 Axon guidance	10.1
hsa04144 Endocytosis	9.0
hsa04310 Wnt signaling pathway	8.6
hsa04530 Tight junction	8.1
hsa04110 Cell cycle	8.0
hsa04350 TGF-beta signaling pathway	6.4
hsa04270 Vascular smooth muscle contraction	6.2
hsa04920 Adipocytokine signaling pathway	6.2
hsa04620 Toll-like receptor signaling pathway	5.0
hsa04370 VEGF signaling pathway	4.8
hsa04810 Regulation of actin cytoskeleton	4.3
hsa04630 Jak-STAT signaling pathway	3.9
hsa04910 Insulin signaling pathway	3.4
hsa04621 NOD-like receptor signaling pathway	3.0
hsa04115 p53 signaling pathway	2.4

3.2 EC-subnetwork genes improve classification accuracy

The ECs themselves are not able to predict recurrence (Fig 3, kappa statics < 0). This is understandable as they are not



(a) Heatmap of TDM scores for DM genes

(b) Heatmap of TDM scores for EC genes

(c) Comparison of patient counts with non-zero TDM scores

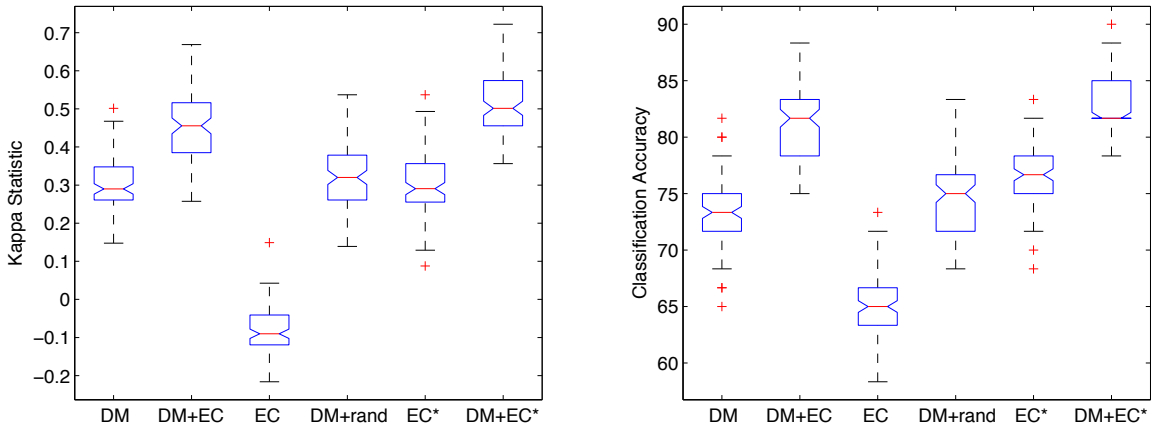
(d) Heatmap of RWB scores for EC genes

Fig. 2. Comparison between DM and EC genes.

differentially methylated, and many ECs have RDM scores only for few patients (Fig 2c). Nevertheless, the combination of ECs and DMs improved the classification accuracy significantly compared to that of DMs alone (0.453 vs. 0.300, kappa statistic, corresponding to 81.2% and 73.4% accuracy, respectively). Note that this improvement cannot be explained by the increased number of features, as combining DMs with 474 randomly seleccted genes only result in minor increase of kappa (0.326) and accuracy (74.4%). Therefore, the improvement of classification accuracy is likely due to the combinatorial effect of DM and EC genes. One possible explaination is that the DMs are universally differentially methylated between recurrent and non-recurrent tumors; while the methylation changes for ECs are patient specific and have weaker statistical significance in terms of differentially methylation. However, the combination of multiple weakly differentiated methylated ECs within the same pathway can be complement to the DM genes in the same pathway and improve classification performance.

3.3 Random walk based scores for EC genes further improve classification accuracy

As many of the metastasis-related genes do not show any differential methylation changes between the recurrent and non-recurrent tumors, we hypothesized that methylation changes may play a role indirectly. One possibility is that the DNA methylation changes of the DM genes may affect the functions of the ECs via protein-protein interactions. Another possibility is the ECs may directly regulate the epigenetic changes of the DMs. In either case, we believe it is possible to measure the relevance between the DMs and ECs based on the network topology. To this end, we used a well-established random walk procedure to compute the probability for a random walker starting at a DM gene to reach any EC genes. The TDM scores of the DM gene is then distributed to the EC genes according to these probabilities. As some of the ECs may also have TDM scores, probabilities were also calculated for a random walker starting at an EC gene to reach other EC genes and the ECs may both distribute its methylation scores to other ECs and receive distributions from other DMs and ECs. These are added together to



(a) Kappa statistic

(b) Classification accuracy

Fig. 3. Classification performance using different types of features. EC*: RWB scores for the ECs are used. rand: randomly selected 474 genes.

derive the inferred methylation impact scores for all the ECs. This procedure is repeated for each patient to obtain a random walk based (RWB) score for each of the EC genes in that patient (see Method). Fig. 5(a) shows the RWB scores of the DMs and ECs on the EC-subnetwork for four selected patients, and Fig. 5(b) shows the impact of these RWB scores on measuring the similarity between patients. Before the random walk, the two recurrent patients had a relatively low similarity compared to the similarity between the non-recurrent and recurrent patients. After the random walk, the similarities between the two recurrent patients and between the two non-recurrent patients both increased, and at the same time the similarities between the recurrent and non-recurrent patients decreased, enabling a perfect separation between recurrence and non-recurrence. Fig. 2(d) shows the RWB scores for all the ECs in all patients. Evidently the score matrix is more dense than the original EC TDM score matrix shown in Fig. 2(b). In fact, although none of the ECs were differentially methylated at p-value < 0.02 according to the TDM scores, with the RWB scores, 203 (43%) of the 474 ECs show statistically significant difference between recurrent and non-recurrent tumors ($p < 0.02$, student's t-test), confirming that the random walk procedure is indeed effective.

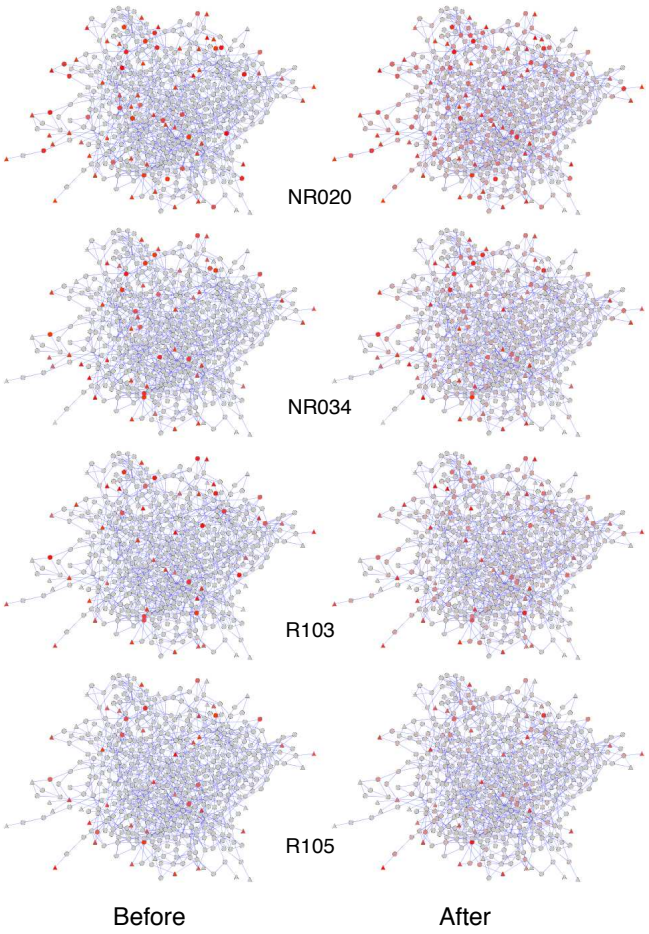
These RWB scores for ECs are then used, either alone or in combination with the TDM scores for the DM genes, to construct a support vector machine classifier to separate recurrent and non-recurrent tumors. As shown in Fig. 3, the performance of the classifier constructed with the RWB scores for the ECs is significantly higher than that of the original TDM scores for the ECs, and even slightly better than that of the DM genes. This is to some extent not surprising, as the RWB scores for the ECs are derived from the TDM scores for the DMs. To see if indeed the ECs provide any additional information other than as proxies to the DM genes, we combined the TDM scores for the DMs and the RWB scores for the ECs. As shown in Fig. 3, this indeed resulted in the highest classification accuracy (kappa statistic 0.513 and accuracy 82.9%). Therefore, it is evident that RWB scores for the ECs provide non-redundant, orthogonal information than the TDM scores of the DM genes.

3.4 Analysis of significant DM and EC markers

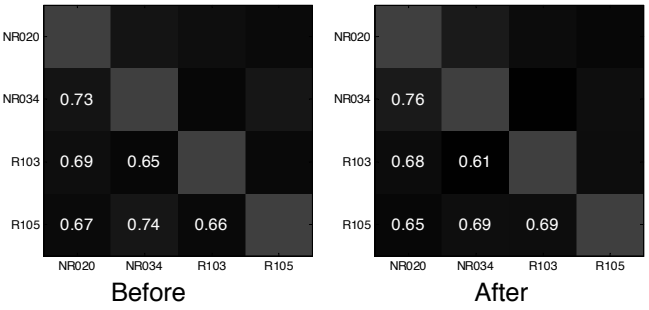
To further understand the role of DM and EC genes as potential biomarkers, we analyzed the normalized feature weights for each of the genes used by the four classifiers based on DM, DM + EC, EC*, and DM + EC*, respectively, where EC* means that the RWB scores rather than the TDM scores are used for the ECs. A larger magnitude of feature weight indicates that the gene is more important for classifying the patients. For genes with positive weights, their hyper-methylation is contributing towards recurrence, and for genes with negative weights their hyper-methylation is contributing towards non-recurrence.

Fig. 5 shows the weights for the genes with a normalized weight $\geq$ 1.5 or $\leq$ -1.5 in at least one classifier. Region I contains DM genes, which may represent universal epigenetic markers. Region II contains EC genes that had non-zero TDM scores in some patients and contributed to the DM + EC classifier; these are EC markers that are epigenetically reprogrammed in specific patients. Finally, Region III contains EC genes that were not used by the DM+EC classifier but had important contributions in the EC* or DM+EC* classifiers. These EC genes themselves are not epigenetically affected; however they are either regulating or regulated by the universal or patient-specific epigenetic markers. As shown in Fig. 5, whenever a feature appears in multiple classifiers, the weights in different classifiers usually have similar sign and magnitude, confirming that the feature weight in support vector machine is a robust measure for the importance of the feature.

Table 2 lists the top ten positively-weighted and negatively-weighted genes from each of the regions described above. While the role of the DM genes in metastasis is not clear, many of the ECs in regions II and III are well known to have important functions in cancer progression and metastasis, such as BRCA1, EPHB2, ID2, ID3, SSTR2, SSTR3, SST, MYOD1, PAX3, HOXD10, and SCT. Interestingly, two genes in region III(a), TLE1 and PARP1, are known as epigenetic regulators (Ali *et al*., 2010; Althaus, 2005; Caiafa *et al*., 2009), confirming our hypotheses that some of the EC markers assume their functions by epigenetically regulating



(a) Patient-specific networks before and after random walk



(b) Similarity between patients before and after random walk

Fig. 4. Comparison between patient-specific networks

the DM genes. Many of the ECs are transcriptional regulators. Their functions may have been regulated by the epigenetic changes occurred in the DM genes whose protein products interact with them.

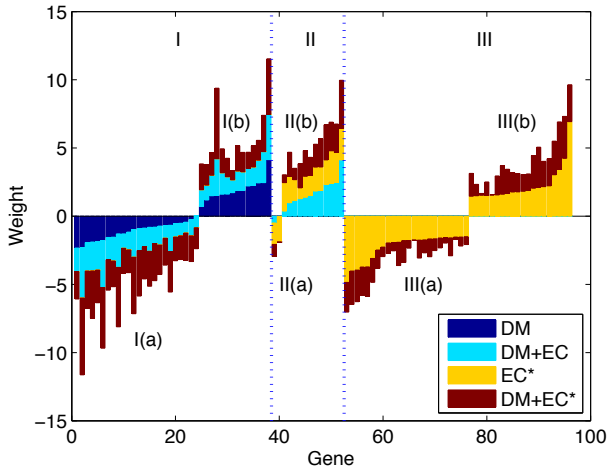


Fig. 5. Feature weights in four SVM-based classifiers for selected genes. Colors depict the classifier from which the weight is obtained. Genes in Region I are DM genes and are sorted by their weights in DM-based SVM classifier; genes in Region II are ECs and are sorted by their weights in DM+EC based classifier; genes in Region III are ECs and are sorted by their weights in EC* based classifier.

Table 2. Top genes from each region in Fig. 5.

I(a)	GALNTL6, FTHL17, ZFP3, SIX6, SPHKAP, POLA1, NPY1R, EPHA5, C19orf34, GABRA2
I(b)	DYNC1I1, CYP26C1, NCRNA00087, ACSS1, TSC22D3, KLHL13, NXPH2, ELK1, TOP3A, ADRA1A
II(a)	SCT, CAMKK1
II(b)	SMC1A, UCN, SST, MYOD1, PAX3, KCNA4, SNCA, TRPC3, HOXD10, EFNA5
III(a)	PARP1, TXNDC17, AVPR1A, SPHK1, TLE1, AES, CORT, PAX6, NFIC, AVPR2
III(b)	EPHB2, COIL, STAU1, GSK3B, ID3, ID2, MCM6, BRCA1, SSTR2, SSTR3

4 CONCLUSIONS

In this paper we have presented a novel network-based algorithm for identifying biomarkers for predicting tumor recurrence from high-throughput DNA methylation data. Our network-based algorithm goes beyond the conventional differential analysis and seek to find both biomarkers with insufficient statistical significance of differential methylation but are within the local neighborhood of the significantly differentially methylated genes, and genes that are not differentially methylated but play important topological roles in connecting the epigenetic markers in the protein-protein interaction networks and therefore are assumed to have functional significance in either regulating or being regulated by the DNA methylation changes of the epigenetic markers. Our results show that the network-based markers are significantly enriched in many KEGG pathways well-known to be involved in tumorgenesis and metastasis, and can be used to significantly improve the

classification accuracy of recurrence, confirming our hypothesis. An unique contribution of this work is that we showed even for the genes without any epigenetic changes, which therefore cannot be considered as epigenetic markers in conventionally analysis, can be utilized to improve the classification performance, suggesting that their functions may have been functionally disturbed by the epigenetic changes of their protein-protein interaction partners.

Our method can be extended in several directions. First, for the EC genes, currently we do not differentiate whether they are regulated by the DM genes, or are regulating the methylation changes of the DMs. This may be partially addressed by including protein-DNA interaction networks where the directions between some nodes can be determined. Second, it is known that markers selected from different datasets for the same disease are usually not comparable. Although our recent results on gene expression data showed that the connectors selected based on Steiner trees are much more robust than the genes selected based on differentially expression (Jahid and Ruan, 2012), it seems that the classification accuracy based on the connectors alone can be further improved. We therefore would like to develop a general classification method that do not depend on the DM genes. For example, after obtaining the EC subnetwork, we may extend the subnetwork to include all genes which are within certain distance to the subnetwork or satisfy some other topological properties. Finally, it may be interesting to combine computational derived subnetworks with subnetworks of signaling pathways that are known to play important roles in specific phenotypes such as metastasis.

ACKNOWLEDGEMENT

Funding: SC3GM086305 (Ruan), U54CA113001 (Huang)

REFERENCES

Ali, S., Zaidi, S., Dobson, J., Shakoori, A., Lian, J., Stein, J., van Wijnen, A., and Stein, G. (2010). Transcriptional corepressor tle1 functions with runx2 in epigenetic repression of ribosomal rna genes. *Proc Natl Acad Sci U S A*, **107**, 4165–9.

Althaus, F. R. (2005). Poly(adp-ribose): a co-regulator of dna methylation? *Oncogene*, **24**(1), 11–12.

Barabasi, A., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat Rev Genet*, **12**, 56–68.

Caiafa, P., Guastafierro, T., and Zampieri, M. (2009). Epigenetics: poly(adp-ribosyl)ation of parp-1 regulates genomic methylation patterns. *The FASEB journal official publication of the Federation of American Societies for Experimental Biology*, **23**(3), 672–678.

Chowdhury, S., Nibbe, R., Chance, M., and Koyuturk, M. (2011). Subnetwork state functions define dysregulated subnetworks in cancer. *Journal of Computational Biology*, **18**, 263–281.

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, **3**, 140–140.

Gatza, M. L., Lucas, J. E., Barry, W. T., Kim, J. W., Wang, Q., Crawford, Datto, Kelley, M., Mathey-Prevot, B., Potti, A., and Nevins, J. R. (2010). A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences*, **107**(15), 6994–6999.

Geistlinger, L., Csaba, G., Kuffner, R., Mulder, N., and Zimmer, R. (2011). From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, **27**, i366–i373.

Goh, K., Cusick, M., Valle, D., Childs, B., Vidal, M., and Barabasi, A. (2007). The human disease network. *Proc Natl Acad Sci U S A*, **104**, 8685–90.

Hidalgo, C., Blumm, N., Barabasi, A., and Christakis, N. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*, **5**, e1000353.

Huang, T. and Esteller, M. (2010). Chromatin remodeling in mammary gland differentiation and breast tumorigenesis. *Cold Spring Harb Perspect Biol*, **2**, a004515.

Huang, Y.-W., Luo, J., Weng, Y.-I., Mutch, D. G., Goodfellow, P. J., Miller, D. S., and Huang, T. H.-M. (2010). Promoter hypermethylation of cidea, haao and rxfp3 associated with microsatellite instability in endometrial carcinomas. *Gynecologic Oncology*, **117**(2), 239 – 247.

Hung, J., Whitfield, T., Yang, T., Hu, Z., Weng, Z., and DeLisi, C. (2010). Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome Biol*, **11**, R23.

Hwang, T., Sicotte, H., Tian, Z., Wu, B., Kocher, J.-P., Wigle, D. A., Kumar, V., and Kuang, R. (2008). Robust and efficient identification of biomarkers by classifying features on graphs. *Bioinformatics*, **24**, 2023–2029.

Jahid, M. J. and Ruan, J. (2012). A randomized steiner tree approach for biomarker discovery and classification of breast cancer metastasis. Submitted to ISMB'12.

Jonsson, P. and Bates, P. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**, 2291–7.

Keller, A., Backes, C., Gerasch, A., Kaufmann, M., Kohlbacher, O., Meese, E., and Lenhof, H. (2009). A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. *Bioinformatics*, **25**, 2787–94.

Kim, Y., Wuchty, S., and Przytycka, T. (2011). Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol*, **7**, e1001095.

Kulis, M. and Esteller, M. (2010). Dna methylation and cancer. *Adv Genet*, **70**, 27–56.

Landis, J. and Koch., G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–74.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008). Inferring Pathway Activity toward Precise Disease Classification. *PLoS Comput Biol*, **4**(11), e1000217+.

Li, J., Lenferink, A., Deng, Y., Collins, C., Cui, Q., Purisima, E., O'Connor-McCourt, M., and Wang, E. (2010). Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun*, **1**, 34.

Liu, M., Liberzon, A., Kong, S., Lai, W., Park, P., Kohane, I., and Kasif, S. (2007). Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet*, **3**, e96.

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S. S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database - 2009 update. *Nucleic Acids Research*, pages 767–772.

Radivojac, P., Peng, K., Clark, W., Peters, B., Mohan, A., Boyle, S., and Mooney, S. (2008). An integrated approach to inferring gene-disease associations in humans. *Proteins*, **72**, 1030–7.

Rauch, T. and Pfeifer, G. (2010). Dna methylation profiling using the methylated-cpg island recovery assay (mira). *Methods*, **52**, 213–7.

Rayward-Smith, V. J. (1983). The computation of nearly minimal Steiner trees in graphs. *Internat J. Math. Ed. Sci. Tech*, **14**, 15–23.

Robinson, M., Stirzaker, C., Statham, A., Coolen, M., Song, J., Nair, S., Strbenac, D., Speed, T., and Clark, S. (2010). Evaluation of affinity-based genome-wide dna methylation data: effects of cpg density, amplification bias, and copy number variation. *Genome Res*, **20**, 1719–29.

Serre, D., Lee, B., and Ting, A. (2010). Mbd-isolated genome sequencing provides a high-throughput and comprehensive survey of dna methylation in the human genome. *Nucleic Acids Res*, **38**, 391–9.

Tong, H., Faloutsos, C., and Pan, J.-Y. (2006). Fast random walk with restart and its applications. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 613–622, Washington, DC, USA. IEEE Computer Society.

Ulitsky, I., Krishnamurthy, A., Karp, R., and Shamir, R. (2010). Degas: de novo discovery of dysregulated pathways in human diseases. *PLoS One*, **5**, e13367.

Vidal, M., Cusick, M., and Barabasi, A. (2011). Interactome networks and human disease. *Cell*, **144**, 986–98.

Vo, S. (1992). Steiner's problem in graphs: heuristic methods. *Discrete Applied Mathematics*, **40**(1), 45 – 72.

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. *Brief Funct Genomics*, pages 2291—-7.

Witten, I. and Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Mateo, CA.

Yildirim, M., Goh, K., Cusick, M., Barabasi, A., and Vidal, M. (2007). Drug-target network. *Nat Biotechnol*, **25**, 1119–26.