Identification of markers associated with global changes in DNA methylation regulation in cancers

Peng Qiu, Li Zhang
Department of Bioinformatics and Computational Biology
University of Texas MD Anderson Cancer Center
Houston, TX, 77030

Abstract—**DNA methylation exhibits different patterns in different cancers. DNA methylation rates at different genomic loci appear to be highly correlated in some samples but not in others. We call such phenomena as conditional concordant relationships (CCRs). In this study, we explored the DNA methylation patterns in 12 common cancers in 2434 patient samples using data collected by The Cancer Genome Atlas (TCGA) project. We developed an exploratory method to characterize CCRs in the methylation data, and identified the 200 most significant gene markers whose on-and-off statuses in DNA methylation are associated with drastic changes in CCRs throughout the genome. Clustering analysis of the methylation data of the 200 markers showed that they are tightly associated with cancer subtypes.**

Index Terms— **TCGA, methylation, conditional correlation.**

I. Introduction

DNA methylation plays an important role in carcinogenesis and progression through hypermethylation to turn off the expression of tumor suppressors and hypothmethylation to activate the expression of oncogenes [1]. Genomic analyses of DNA methylation using microarrays and next generation sequencing technologies revealed that various forms of neoplasia and cancers are associated with massive changes in DNA methylation [2]. Such changes are often distinctive depending on the subtypes of cancers [3]. DNA methylation in the cells is apparently regulated by a large intricate network. However, while a large number of genomic network studies focused on data of gene expression, protein-protein interactions, and protein-DNA/RNA interactions [4], [5], little has been done to incorporate DNA-methylation data to understand the underlying regulatory network.

In general, relationships that link different genes at DNA, RNA, protein, and metabolites levels strongly depend on the specific context, such as cell type, sub-cellular location and time of the biological processes. A number of methods have been developed to uncover context-dependent relationships using gene expression data. For example, the liquid association model was developed to identify mediator genes that can modulate coexpression of other pairs of genes [6]. A few other similar models were also proposed to describe three-way relationships among genes [7]–[9]. Cancer type dependent coexpression patterns were reported in [10], [11]. In [12], the MINDy algorithm used conditional mutual information to identify modulators that strongly affect the concerted activities of transcription factors and their targets, and found novel modulators of MYC function in B cells.

In this study, we focused on the dynamic nature of concordant relationships between the methylation status of genes, using a large DNA methylation dataset of 2434 samples across 12 cancer types generated by The Cancer Genome Atlas (TCGA) project. We observed that many gene-pairs showed dramatic changes in methylation pattern in different cancers. We call such phenomena as conditional concordant relationships (CCRs). We are interested in finding marker genes that have the following property: depending on the methylation status of the marker, the patient samples can be dichotomized into two groups, and the gene-gene correlation matrices derived from the methylation data of the two groups are drastically different. Such markers are likely to be associated with global changes in the methylation correlation patterns. The concept of the methylation markers resembles the modulator in the three-way gene expression studies [7]. We developed a method to identify such markers, and demonstrated the utility of our approach to study CCRs, classify cancer subtypes, and explore the patterns of DNA methylation in cancer.

II. Results

A. Genomic patterns of DNA methylation in cancers

To show the overall pattern of DNA methylation in cancers, we downloaded methylation data of 2434 samples across 12 cancer types from the TCGA data portal [13], and performed hierarchical clustering analysis. Table I showed the sample size of each of the 12 cancer types. This dataset contains 27,578 probes interrogating proximal promoter regions of 14,475 genes in the human genome. The methylation status of many probes showed small variance, and therefore do not contribute to the clustering analysis. We hence removed the non-changing probes and kept $\sim$9000 probes that have the highest variance across samples. We also removed probes on the X and Y chromosomes because their methylation rates mainly reflected gender difference rather than disease or tissue differences. Figure 1 showed the cluster diagram generated from the 9000 probes across 2434 samples, with each row representing one probe and each column representing one sample. The bottom panel showed the tissue type and normal-cancer status of the samples. It can be observed that the samples were mostly organized by tissue types, with some noticeable outliers. GBM, LAML, OV, BRAC and UCEC samples formed their own clusters. READ and COAD samples were grouped together. The kidney cancer samples and the normal kidney samples

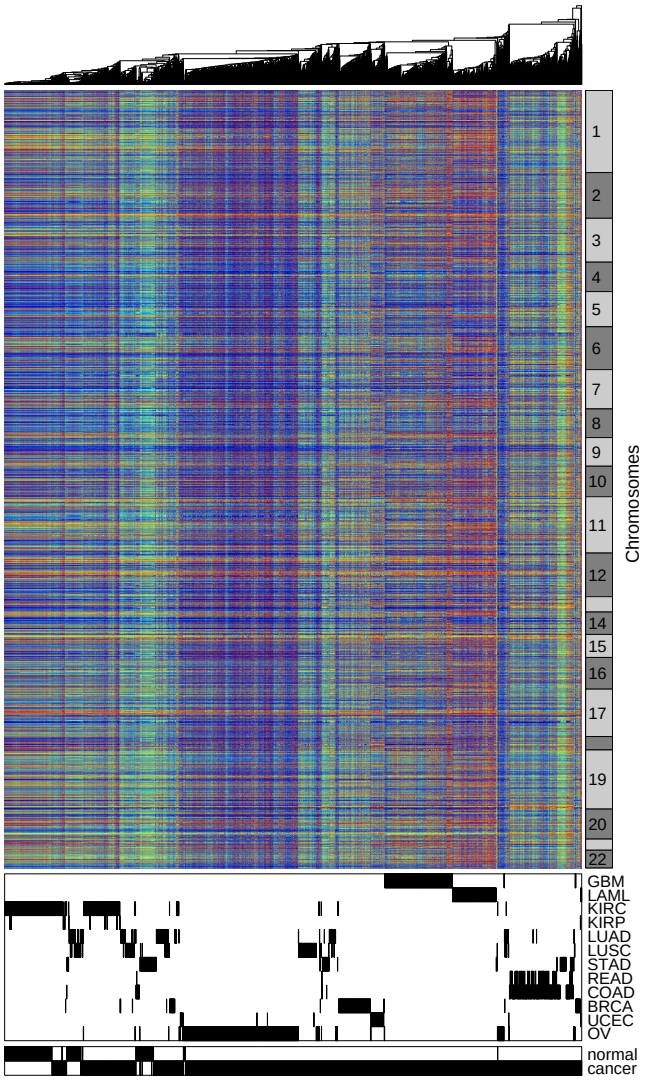

Fig. 1. Clustering diagram of the whole genomic pattern observed in 2434 samples across 12 common cancers.

TABLE I
CANCER TYPE AND SAMPLE SIZE OF TCGA METHYLATION DATA

Cancer Type	Sample size	
	cancer	normal
GBM - Glioblastoma multiforme	291	0
LAML - Acute Myeloid Leukemia	188	0
KIRC - Kidney renal clear cell carcinoma	219	199
KIRP - Kidney renal papillary cell carcinoma	16	5
LUAD - Lung adenocarcinoma	128	24
LUSC - Lung squamous cell carcinoma	134	27
STAD - Stomach adenocarcinoma	82	59
READ - Rectum adenocarcinoma	70	1
COAD - Colon adenocarcinoma	168	15
BRCA - Breast invasive carcinoma	186	0
UCEC - Uterine Corpus Endometrioid Carcinoma	70	0
OV - Ovarian serous cystadenocarcinoma	542	10

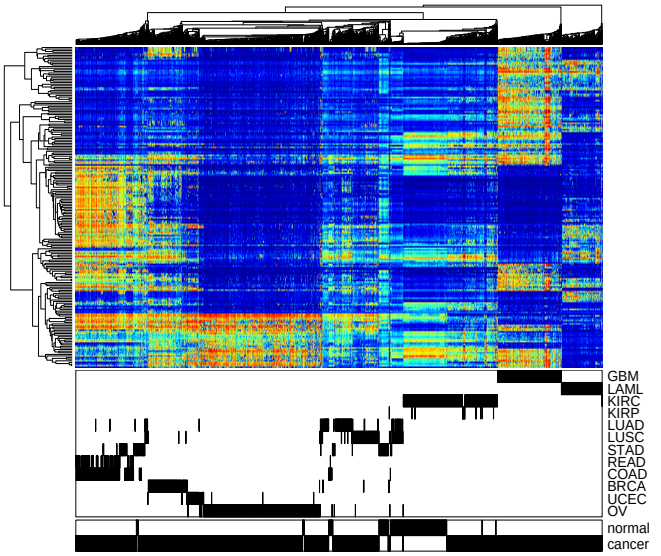

Fig. 2. Clustering diagram of 2434 cancer samples and 200 top probes associated with CCRs.

were clustered close to each other. The normal and cancer samples of LUAD and LUSC were mixed with each other, but scattered across the clustering diagram. The majority of the lung samples appeared to be similar to KIRC. The STAD samples formed three groups. A subset of the STAD cancer samples were clustered with READ and COAD, while the remaining STAD cancer samples were clustered with lung cancer samples. The STAD normal samples appeared to be similar to another group of lung cancer samples.

B. Identifying the markers associated with global changes in gene-gene correlations

To systematically evaluate the CCRs, we searched for marker probes associated with a large number of CCRs. We randomly selected 1500 samples as the training set, and the remaining samples were reserved as the testing set. Based on the training samples, we selected $\sim$9000 high variance probes, derived scores to evaluate each probe's association with CCRs, and rank-ordered the probes (see Methods). These high variance probes were also scored based on the testing set. The scores derived from training and testing data were highly correlated (Pearson correlation > 0.95), suggesting that the top ranked probes and their scores were robust.

We performed clustering analysis of all 2434 samples based on the top 200 probes selected from the training set. As shown in Figure 2, the top 200 CCR-associated probes were able to separate cancer types. Similar to the previous analysis based on 9000 high variance probes, the top CCR-associated probes defined distinct clusters for GBM, LAML, OV, BRAC and UCEC, respectively. READ and COAD samples were grouped into one cluster. The major difference was the clustering of the lung samples. Based on the CCR-associated probes, the two subtypes of lung samples (LUAD and LUSC) formed one tight cluster. Normal lung samples were grouped with KIRCs in the previous analysis, but the CCR-associated probes highlighted the difference between them.

For most cancer types where normal and cancerous samples were both available, the cancerous and corresponding normal

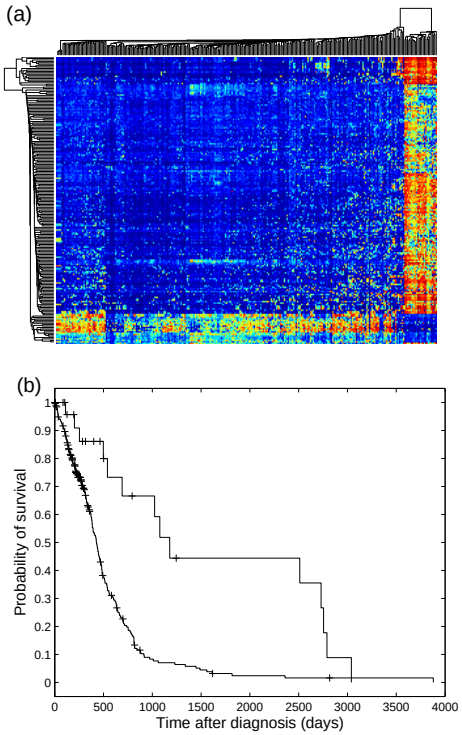

Fig. 3. (a) Clustering and diagram of 291 GBM samples based on top 200 CCR-associated probes. (b) Kaplan-Meier plot of the survival data of the two GBM subgroups observed in (a). The survival of the two groups showed significant differences (logrank test pvalue 10^{-5}).

samples were clustered close to each other. This observation suggests that the methylation difference across different tissue types is larger than cancer-induced methylation changes. The only exception in this dataset was STAD. In Figure 2, we observed that the STAD normal samples were more similar to the lung samples, while the STAD cancer samples were more similar to the COAD and READ samples. This observation indicated that methylation might play a major role in stomach adenocarcinoma.

C. CCR-associated markers recovered GBM subtypes

In the previous section, we showed that when applied to all samples containing multiple tissue types, the top CCR-associated markers were able to separate tissue types. A natural next step was to focus on one cancer type, and examine whether the CCR-associated markers can identify cancer subtypes. We focused on the 291 GBM samples, selected ∼9000 high variance probes, scored each probe's association with CCRs, rank-ordered the probes, and used the top 200 probes to perform clustering analysis. Figure 3 (a) showed the clustering diagram of the 291 GBM samples based on the top 200 CCR-associated probes. We observed that the GBM samples were divided into two groups. The clinical outcome of the smaller group was significantly better than the bigger group, as shown in Figure 3 (b). The smaller GBM sample group with better survival was first reported in [14]. This group of samples carry a CpG island methylator phenotype, which is associated with better survival and low-grade gliomas. In [14], clustering analysis was performed based on 1500 high variance probes, and discovered three GBM subtypes. One of the three was the smaller sample group in Figure 3 (a). The remaining two corresponded to the bigger group in our analysis, but there was not significant evidence for the biological and clinical difference between them.

III. Discussion

We described an approach to explore complex patterns observed in DNA methylation data. We identified conditional concordant relationships (CCRs) and markers associated with global changes of methylation correlation in different cancers. Expectedly, when the identified markers were used for clustering analysis, the clustering diagram largely coincided with cancer types, since distinct methylation patterns exist in different tissue types. We demonstrated that our approach can be used to uncover tissue types and subtypes of cancer. In this sense, our method is similar to feature selection and unsupervised clustering.

The current study is limited to methylation data only. However, data from multiple platforms measuring gene expression, microRNA expression, DNA copy number and somatic mutations can all be evaluated as candidate markers that affect CCRs in DNA methylation. Integrating data from multiple platforms will be increasingly powerful as more data are being accumulated in the TCGA project.

IV. Method

A. Data and Preprocessing

In this study, we focused on the DNA methylation data provided by The Cancer Genome Atlas (TCGA) project. Genome-wide methylation measurements of 2434 samples were available, spanning across 12 cancer types. The data were generated using the IIllumina Infinium Human DNA Methylation27 array platform, which interrogates the methylation status of 27,578 CpG sites for each sample.

We used the level 3 methylation data defined by TCGA, which is the ratio of $M_i/(U_i + M_i)$ for each CpG site i. M_i represents the methylated probe intensity of CpG site i, while U_i is the unmethylated probe intensity. Therefore, the numerical range of the data is between 0 and 1. 0 means unmethylated, and 1 means completely methylated. The data contains null entries, which correspond to probes that overlap with known single nucleotide polymorphisms (SNPs) or other genomic variations, and probes whose signal intensities are lower than the background. In our analysis, we filtered out probes with many null entries (number of nulls more than 1% of the sample size) and probes with small standard deviation ($SD < 0.1$). Roughly 9000 probes survived these two filtering criteria, and were considered in the analysis of conditional concordant relationships.

B. Dichotomize samples based on methylation

Although DNA methylation is a reversible process and methylated CpG sites may not be completely methylated,

methylation data appear to be bimodal in general. By thresholding the ratio $M_i/(U_i + M_i)$ (i.e. nominal threshold 0.2), we can use probe i to divide samples into two groups. The status of CpG site i in one group is unmethylated, whereas the CpG site i in the other group is methylated. If the methylation correlation patterns in the two sample groups are quite different, the CpG site i is likely to be related to the global changes of methylation regulation.

C. Clustering

Before calculating the changes of methylation correlation, clustering is performed to find modules of highly correlated probes. The purpose is to reduce the computational complexity. The pairwise correlations between the modules can be used as surrogates of the pairwise correlations between individual probes.

We use a variation of agglomerative clustering algorithm [15], [16]. This algorithm requires a user-specified threshold for cluster coherence, defined as the average Pearson correlation between each probe in the cluster and the cluster mean. This parameter determines the quality of the resulting clusters (the default setting is 0.7). At the beginning of the first iteration of the agglomerative algorithm, each probe forms its own cluster. One probe is randomly chosen and merged with its nearest neighbor defined by Pearson correlation and average linkage, and these two probes become unavailable in the current iteration. Then, another probe is randomly chosen from the remaining ones and merged with its nearest neighbor, if the nearest neighbor is still available. Again, the chosen probe and its nearest neighbor become unavailable in the current iteration. If a merge results in a cluster whose coherence is below the user-specified threshold, the merge is rejected. After all the probes become unavailable, the first iteration ends and the number of clusters is reduced by approximately half. The same procedure is repeated in the second iteration to further reduce the number of clusters. The iterative process continues until all merges in a particular iteration are rejected.

The algorithm guarantees that the quality of all the resulting probe clusters is higher than the user-specified threshold. The average of each cluster can be viewed as a meta-probe that summarizes the average methylation status of the cluster of correlated probes.

D. Identify CCR-associated switch-like probes

To identify CCR-associated probes, we used the training samples to filter for roughly 9000 probes that had small number of null entries and high standard deviation. These probes were considered as candidates to be evaluated. We also performed an agglomerative algorithm using the training set, to cluster probes into modules that contained highly correlated probes, and represented each module by the mean methylation profile of probes in that module.

For each candidate probe, we evaluated whether its on-off status affects methylation correlation globally. We dichotomized the training samples into two groups (i.e. threshold = 0.2), computed the module-module correlation matrices for the two sample groups separately, performed z-transform, and summarized the difference between the two correlation matrices into one scalar score ($s = \sum_{i,j} |z_1(i,j) - z_2(i,j)|$). If a candidate probe resulted in an extremely unbalanced split (i.e. the smaller sample group contained less than 15% of the samples), this candidate probe was not scored, because correlation based on small number of samples may not be accurate and reliable. The candidate probes were rank-ordered according to their scores, where the methylation status of top ranking probes were associated to large changes of methylation correlation.

REFERENCES

[1] E. Ballestar, "An introduction to epigenetics", *Adv Exp Med Biol.*, vol. 711, pp. 1–11, 2011.

[2] P. A. Jones, S. B. Baylin, "The fundamental role of epigenetic events in cancer", *Nat Rev Genet.*, vol. 3, no. 6, pp. 415-428, 2002.

[3] N. G. Bediaga, A. Acha-Sagredo, I. Guerra, A. Viguri, C. Albaina, I. Ruiz Diaz, R. Rezola, M. J. Alberdi, J. Dopazo, D. Montaner, M. de Renobales, A. F. Fernndez, J. K. Field, M. F. Fraga, T. Liloglou, M. de Pancorbo, "DNA methylation epigenotypes in breast cancer molecular subtypes", *Breast Cancer Res.*, vol. 12, no. 5, pp. R77, 2010.

[4] G. Karlebach, R. Shamir, "Modelling and analysis of gene regulatory networks", *Nat Rev Mol Cell Biol.*, vol. 9, no. 10, pp. 770-780, 2008.

[5] T. M. Przytycka, M. Singh, D. K. Slonim, "Toward the dynamic interactome: it's about time.", *Brief Bioinform.*, vol. 11, no. 1, pp. 15-29, 2010.

[6] K. C. Li, C. T. Liu, W. Sun, S. Yuan, T. Yu, "A system for enhancing genome-wide coexpression dynamics study", *Proc Natl Acad Sci U S A.*, vol. 101, no. 44, pp. 15561-15566, 2004.

[7] J. Zhang, Y. Ji, L. Zhang, "Extracting three-way gene interactions from microarray data", *Bioinformatics*, vol. 23, no. 21, pp. 2903-2909, 2007.

[8] M. Kayano, I. Takigawa, M. Shiga, K. Tsuda, H. Mamitsuka, "Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data", *Bioinformatics*, vol. 25, no. 21, pp. 2735-2743, 2009.

[9] Y. Y. Ho, G. Parmigiani, T. A. Louis, L. M. Cope, "Modeling liquid association", *Biometrics*, vol. 67, no. 1, pp. 133-141, 2011.

[10] J. K. Choi, U. Yu, O. J. Yoo, S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer", *Bioinformatics*, vol. 21, no. 24, pp. 4348-4355, 2005.

[11] M. Dettling, E. Gabrielson, G. Parmigiani, "Searching for differentially expressed gene combinations", *Genome Biol.*, vol. 6, no. 10, pp. R88, 2005.

[12] K. Wang, M. Saito, B. C. Bisikirska, M. J. Alvarez, W. K. Lim, P. Rajbhandari, Q. Shen, I. Nemenman, K. Basso, A. A. Margolin, U. Klein, R. Dalla-Favera, A. Califano, "Genome-wide identification of post-translational modulators of transcription factor activity in human B cells", *Nat Biotechnol.*, vol. 27, no. 9, pp. 829-839, 2009.

[13] Cancer Genome Atlas Research Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways", *Nature*, vol. 455, no. 7216, pp. 1061-1068, 2008.

[14] H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat, R. G. Verhaak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, K. Aldape, Cancer Genome Atlas Research Network, "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma", *Cancer Cell*, vol. 17, no. 5, pp. 510-522, 2010.

[15] P. Qiu, A. J. Gentles, S. K. Plevritis, "Discovering biological progression underlying microarray samples", *PLoS Comput Biol.*, vol. 7, no. 4, pp. e1001123, 2011.

[16] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs Jr., R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, S. K. Plevritis, "Extracting a Cellular Hierarchy from High-dimensional Cytometry Data with SPADE", *Nature Biotechnology*, in press, 2011.