2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)
December 2-4, 2012, Washington, DC, USA

Integrative analysis of methylation and gene expression data in TCGA

Yihua Liu and Peng Qiu*
Department of Bioinformatics and Computational Biology
The University of Texas MD Anderson Cancer Center
Houston, Texas 77030
Email: yliu13@mdanderson.org, pqiu@mdanderson.org

***Abstract*—An integrative analysis of methylation and gene expression is performed using TCGA data for 997 samples from multiple cancer types. We use conditional mutual information to examine each gene separately, identify 798 genes whose expression are repressed by their methylation, and derive gene-specific thresholds to dichotomize their methylation data. For each methylation controlled gene, we evaluate its conditional correlation with other genes, and infer which other genes are its regulators or targets. The resulting set of regulator genes is much larger than that of target genes, and the set of regulator genes is highly enriched by transcription factors.**

I. Introduction

DNA methylation plays an important role in cancer through hypermethylation to turn off tumor suppressors [1], [2] and hypomethylation to activate oncogenes [3]. It is widely recognized that DNA methylation is associated with silencing of gene expression [4]. With data from high-throughput array and sequencing technologies, a number of studies have reported integrative analysis of methylation and gene expression [5]–[7].

Integrating methylation and gene expression data can lead to better understanding of regulatory relationships among genes. In the literature, great efforts have been made to study gene regulatory networks using expression data. Pairwise correlation and mutual information have been used to construct gene networks [8]. In addition, several methods for uncovering context-dependent relationships have been proposed. For example, the liquid association model was developed to identify genes that can modulate coexpression of other pairs of genes [9]. A few other similar models have been proposed to describe three-way relationships among genes [10], [11]. Conditional mutual information has been used to identify regulators that strongly affect the concerted activities of transcription factors and their targets [12]. These methods can be applied to integrate methylation and gene expression data, uncovering how relationships among genes are dependent on the epigenetic context of DNA methylation.

To integrate methylation and gene expression, an ideal resource is a large collection of samples for which both data are available. The Cancer Genome Atlas (TCGA) project provides such data [13]. Moreover, the TCGA samples are derived from multiple cancer and tissue types. The diversity among the samples may enable us to see relationships that cannot be observed in individual tissue types.

In this paper, we analyze DNA methylation and gene expression data in TCGA. Conditional mutual information is used to examine the relationship between the methylation status of a gene and its expression, to identify which genes are significantly repressed by their methylation and derive gene-specific methylation thresholds. For each methylation controlled gene, we use conditional correlation to examine the three-way relationships among the methylation and expression of that gene, and the expression of another gene, for the purpose of inferring directional regulatory relationships among genes.

II. TCGA data and preprocessing

We use DNA methylation data and gene expression data provided by TCGA. Genome-wide methylation measurements were generated using the Illumina Infinium Human DNA Methylation27 array platform, which interrogates the methylation status of 27,578 CpG sites in proximal promoter regions of 14,475 genes in the human genome. As of February 25, 2012, methylation data for 3382 samples across 12 cancer types were available. We downloaded the TCGA level 3 preprocessed data, which is the ratio $M_i/(U_i + M_i)$ for each CpG site i. M_i represents the intensity of the methylated probe for CpG site i, whereas U_i is the unmethylated probe intensity. Therefore, the numerical range of the data is between 0 and 1. 0 indicates unmethylated, whereas 1 indicates completely methylated. The data contains a small fraction of empty entries, because the corresponding probes either overlap with known single nucleotide polymorphisms (SNPs) or other genomic variations, or their signal intensities are lower than the background.

TCGA uses several platforms to quantify gene expression, among which the Illumina GA II and HiSeq platforms profiled the largest number of samples. As of February 25, 2012, RNA-seq data for 2271 samples across 9 cancer types were available. Again, we downloaded preprocessed data, which are the RPKM values for 20532 genes in each sample. The numerical range of the data is between 0 and 10^5. For each gene, we replace the zero entries with the minimal non-zero

978-1-4673-5236-9/12/$31.00 ©2012 IEEE

value of this gene across all samples, and transform the data to log2 scale.

The number of overlapping samples between the methylation data and the gene expression data is 997, which covers 7 different cancer types. The three most prevalent cancer types among those samples are breast invasive carcinoma (306 samples), kidney renal clear cell carcinoma (207 samples) and colon adenocarcinoma (161 samples). Our integrative analysis is performed based on these 997 overlapping samples.

III. Gene-specific methylation on-off threshold

Methylation is often described as a binary on-off signal [14], and it is widely recognized that methylation represses gene expression. Typically, if a gene is controlled by its methylation, its expression is low when methylated. On the other hand, when unmethylated, its expression can be either high or low. Since measurements for methylation and expression are both continuous, a biaxial plot of these two signals will exhibit an L-shape pattern, similar to those shown in Figure 1. If we truly believe that methylation is binary, there are two implications: (1) the reflection point of the L-shape is an appropriate choice to binarize methylation data, and (2) conditioning on the binarized on-off methylation status, the continuous valued methylation data and expression data should be independent, which motivates us to quantify the L-shape pattern using conditional mutual information (MI) [15]. In this section, we use TCGA data to ask two questions: which genes exhibit L-shape, and what is the optimal threshold for binarizing methylation data for each L-shape gene.

To determine whether methylation and expression of a gene exhibit an L-shape, we compute the conditional MI for different choices of thresholds to binarize the methylation data. If we consider the continuous valued methylation and expression data as two random variables X and Y, and denote a nominal threshold as t, the conditional MI can be written as a weighted sum of MIs on the two sides of the threshold.

$$cMI(t) = I(X,Y|X<t)p(X<t)+I(X,Y|X\geq t)p(X\geq t) \tag{1}$$

When t is chosen to be 0 or 1, all data points are on one side of the threshold, and the conditional MI equals to the mutual information derived from all data points. For an L-shape gene, as t moves from 0 to 1, $cMI(t)$ first decreases and then increases, and its value approaches zero when t coincides with the reflection point (Figure 1). Therefore, the ratio $r = \frac{min(cMI(t))}{cMI(0)}$ for an L-shape gene is small, and $t^* = argmin(cMI(t))$ is the optimal threshold for dichotomizing the methylation data of this gene. Although some other patterns may also lead to small r, this ratio can serve as a criterion to filter out genes that do not exhibit L-shape.

To estimate the MI terms in equation (1) from the TCGA data, we use a kernel-based estimator [16], which constructs a joint probability distribution by applying a Gaussian kernel to each data point, and estimates the MI based on the joint distribution. The estimator is as follows:

$$I(X;Y) = \frac{1}{M}\sum_i log\frac{M\sum_j e^{-\frac{1}{2h^2}((x_i-x_j)^2+(y_i-y_j)^2)}}{\sum_j e^{-\frac{1}{2h^2}(x_i-x_j)^2}\sum_j e^{-\frac{1}{2h^2}(y_i-y_j)^2}} \tag{2}$$

where h is a tuning parameter for the kernel width, i and j are indices for samples. In our analysis, we normalize the methylation and expression data to zero-mean-unit-variance and empirically set $h = 0.3$ [16].

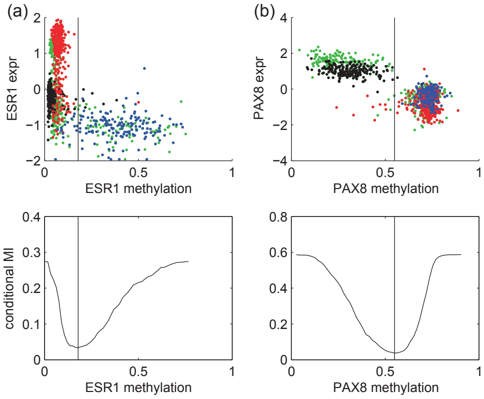

Fig. 1. Two example genes whose expression is controlled by methylation, panel (a) corresponds to ESR1 and panel (b) corresponds to PAX8. In both examples, the top panels show the scatter plots of methylation and expression data. Colors are used to distinguish cancer types: breast cancer samples are in red; kidney samples are in black; colon samples are in blue; and green represents the other cancer type samples. The bottom panels show the conditional MI for different choices of methylation thresholds. Appropriate thresholds can be suggested by the minimum of conditional MI.

For each CpG site, we examined the conditional MI between its methylation and the expression of its corresponding gene. Then, we filtered for L-shapes using a combination of three criteria: the ratio $r < 0.25$; unconditioned MI $cMI(0) > 0.1$; the average expression on the left side of the optimal threshold t^* is higher than the average expression on the right side. The parameters here are chosen according to a random permutation test and p-value of 0.01. According to the above criteria, a total of 798 genes are selected to be L-shape genes. Two examples are shown in Figure 1. For ESR1, the optimal threshold to binarize its methylation is 0.18, whereas the optimal threshold for PAX8 is 0.55. This result suggests that the optimal thresholds can be quite different for different genes. Indeed, if we plot the histogram of the identified thresholds for all the 798 L-shape genes, we see a wide-spread distribution in Figure 2, indicating that the optimal threshold for dichotomizing methylation data is highly gene-specific.

In both examples shown in Figure 1, the colors encode for different disease types. Red, black and blue represent breast, kidney and colon cancers, respectively. Green represents all other cancer types in the data. When data for all disease types are visualized together, it is easy to observe the L-shape relationship. However, since dots with the same color tend to appear on the same side of the optimal threshold, if we focus our attention on only one disease type, we will not be able

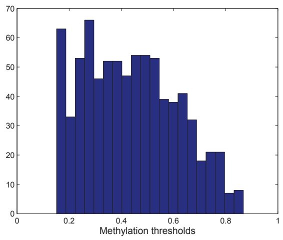

Fig. 2. Histogram of gene-specific methylation thresholds for the 798 identified genes whose expression are controlled by their methylation.

to observe the L shape. This shows the power of analyzing multiple cancer types together.

IV. Inferring directional gene regulatory relationships

In the previous section, we identify L-shape genes whose expressions are controlled by methylation and their gene-specific methylation on-off thresholds. For each L-shape gene i, we can use its specific methylation threshold to divide samples into two groups, one with gene i unmethylated and the other with gene i methylated. This enables calculation of the conditional correlations between gene i and other genes, conditioning on the methylation status of i. The conditional correlations may be used to infer directional regulatory relationships between i and other genes.

More specifically, we are interested in gene pairs that exhibit patterns shown in Figure 3. If expressions of i and j are highly correlated when i is unmethylated, and j's expression is tightly constrained when i's expression is controlled by its methylation, we will observe a pattern similar to Figure 3 (a). Such a pattern indicates that expression of gene j is highly correlated with gene i regardless of i's methylation status, which implies that gene j is likely to be a target of gene i. On the other hand, if j's expression is not constrained when i is methylated, the data will exhibit a pattern similar to Figure 3 (b), which implies that gene j is likely to be a regulator of gene i, and j only regulates i when i is not silenced by its own methyaltion. Unfortunately, the pattern in Figure 3 (b) does not rule out a more complex relationship in which gene i and j are both regulated by some other gene.

To search for patterns similar to Figure 3, we examine the conditional correlations between each of the 798 methylation controlled genes and all other genes. For gene j to be considered a target of a methylation controlled genes i, we require: absolute value of their overall correlation $|corr(i,j)| > 0.6$; when i is unmethylated, $|corr(i,j|i_{meth} \text{ off})| > 0.6$ and j's expression range is larger than 60% of its overall range; when i is methylated, $|corr(i,j|i_{meth} \text{ on})| < 0.4$ and j's expression range is smaller than 40% of its overall range. Using this set of criteria, 531 gene pairs are identified, corresponding to 112 unique target genes. One examples is shown in Figure 4 (a). The left panel shows that GUCY2C's expression is repressed by its methylation. In the middle panel, we observe a positive correlation between GUCY2C and NOX1 in samples whose GUCY2C is unmethylated. The right panel shows that NOX1's expression is tightly constrained when GUCY2C is methylated. This combination indicates that NOX1 is likely to be a target activated by GUCY2C. Another example is shown in Figure 4 (b), suggesting that PKP3 is a target which is inhibited by EDNRB.

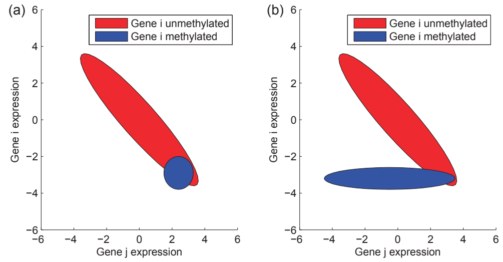

Fig. 3. Possible patterns of expressions of i and j conditioning on the methyaltion status of i. Such pattern can be used to infer directional regulatory relationships, i.e. (a) indicates j is a target of i, whereas (b) indicates j is a regulator of i.

Another set of criteria is used to identify genes that are potentially regulators of the methylation controlled genes: when i is unmethylated, $|corr(i,j|i_{meth} \text{ off})| > 0.6$ and j's expression range is larger than 60% of its overall range; when i is methylated, $|corr(i,j|i_{meth} \text{ on})| < 0.4$ but j's expression range is still larger than 60% of its overall range. This set of criteria generates 21911 gene pairs, corresponding to 8057 unique genes regulating the methylation controlled genes. Two examples are shown in Figure 5, suggesting that MYB regulates and activates LGALS4, and EME1 is likely to regulate and inhibit SLC3A1.

To investigate the validity of the directional relationships, we compare the identified sets of regulator genes and target genes against a manually curated list of 1391 transcription factors (TFs) [17]. Using the hypergeometric test, we notice significant overlap between the 8057 regulator genes and the TFs ($pvalue = 0.00013$). In contrast, the overlap between the 112 target genes and the TFs is insignificant (hypergeometric test $pvalue = 0.74$). This comparison supports our results, because if the identified directional relationships are valid, we would expect that the set of regulator genes is more enriched by TFs than the set of target genes.

V. Discussion

We perform integrative analysis of methylation and gene expression data in TCGA. Using the conditional mutual information, we identify 798 genes whose expression are regulated by their methylation status, and derive gene-specific thresholds to determine the on-off methylation status for those genes. We further compare the conditional correlation between the methylation controlled genes and others to infer directional regulatory relationships, i.e., which genes are regulators or targets of the methylation controlled genes. The number of the identified regulator genes is much larger than that of the target genes, and the regulator genes are much more highly enriched by TFs.

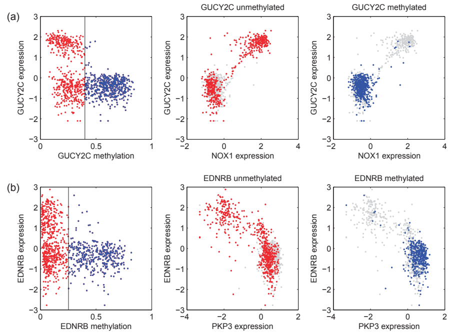

Fig. 4. Examples of identified target genes of methylation controlled genes. (a) Relationship between GUCY2C (methylation controlled gene) and NOX1. The red, blue and gray dots collectively represent all 997 samples. Red highlights samples with GUCY2C unmethylated, and blue highlights samples whose GUCY2C is methylated. This pattern indicates NOX1 is a target activated by GUCY2C. (b) Methylation and expression of EDNRB and expression of PKP3, indicating that PKP3 is a target which is inhibited by EDNRB.

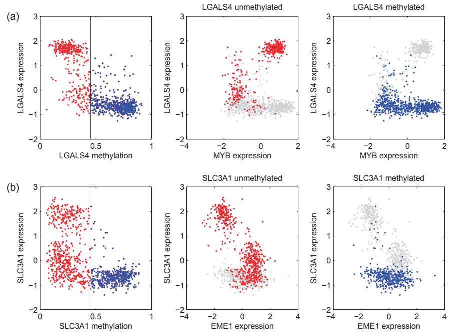

Fig. 5. Examples of potential regulator genes of methylation controlled genes. (a) Methylation and expression of LGALS4 and expression of MYB, indicating that MYB regulates and activates LGALS4. (b) Methylation and expression of SLC3A1 and expression of EME1, indicating that EME1 regulates and inhibits SLC3A1.

For future work, there are a few possible extensions. The methylation data used here are generated by the Illumina Methylation 27k platform, which measures 27,578 CpG sites annotated to 14,475 genes. TCGA also generates methylation data using another platform, the Illumina Methylation 450k array, which measures roughly 20 times more CpG sites. We plan to redo the analysis using the 450k methylation data, which will enable us to identify more methylation controlled genes and a larger number of directional regulatory relationships. Another extension is to assemble the identified directional relationships into a directed gene network, and study the biological relevance of hub nodes and root nodes of the network.

In this analysis, samples from multiple cancer and tissue types are included. The diversity in such a pan-cancer dataset empowers the analysis to identify more relationships. For example, many of the L-shape patterns between methylation and gene expression can only be observed when multiple tissue types are analyzed together. However, such a dataset will also lead to many false positives that reflect tissue differences rather than regulatory mechanisms. Therefore, there is a dilemma of what data to use, multiple cancer types together or individual diseases separately. To better understand this question, we will repeat our analysis in individual cancer types separately, and compare the results with that observed here.

Acknowledgment

This work is partially supported by TCGA Genome Data Analysis Center grant at UT MD Anderson Cancer Center U24 CA143883 02 S1, and NIH grant R01 CA163481.

References

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

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

[3] M. Widschwendter, G. Jiang, C. Woods, H. Mller, H. Fiegl, G. Goebel, C. Marth, E. Mller-Holzner, A. Zeimet, P. Laird, and M. Ehrlich, "DNA hypomethylation and ovarian cancer biology," *Cancer Res.*, vol. 64, no. 13, pp. 4472 – 4480, 2004.

[4] J. Newell-Price, A. Clark, and P. King, "DNA methylation and silencing of gene expression," *Trends Endocrinol Metab.*, vol. 11, no. 4, pp. 142 – 148, 2000.

[5] M. Li, C. Balch, J. Montgomery, M. Jeong, J. Chung, P. Yan, T. Huang, S. Kim, and K. Nephew, "Integrated analysis of dna methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer," *BMC Med Genomics*, vol. 2, p. 34, 2009.

[6] R. Shaknovich, H. Geng, N. Johnson, L. Tsikitas, L. Cerchietti, J. Greally, R. Gascoyne, O. Elemento, and A. Melnick, "DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma," *Blood*, vol. 116, no. 20, pp. e81–89, 2010.

[7] The Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[8] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis." *Stat Appl Genet Mol Biol*, vol. 4, no. 1, 2005.

[9] K. Li, C. Liu, W. Sun, S. Yuan, and T. Yu, "A system for enhancing genome-wide coexpression dynamics study," *Proc Natl Acad Sci U S A.*, vol. 101, no. 44, pp. 15 561–15 566, 2004.

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

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

[12] K. Wang, M. Saito, B. Bisikirska, M. Alvarez, W. Lim, P. Rajbhandari, Q. Shen, I. Nemenman, K. Basso, A. Margolin, U. Klein, R. Dalla-Favera, and 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] The Cancer Genome Atlas Research Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330–333, 2012.

[14] P. Qiu and L. Zhang, "Identification of markers associated with global changes in DNA methylation regulation in cancers," *BMC Bioinformatics*, vol. 13, no. Suppl 13, p. S7, 2012.

[15] T. Cover and J. Thomas, *Elements of Information Theory 2nd Edition*, ser. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006.

[16] P. Qiu, A. J. Gentles, and S. K. Plevritis, "Fast calculation of pairwise mutual information for gene regulatory network reconstruction," *Computer methods and programs in biomedicine*, vol. 94, pp. 177–180, 2009.

[17] J. Vaquerizas, S. Kummerfeld, S. Teichmann, and N. Luscombe, "A census of human transcription factors: function, expression and evolution," *Nature Reviews Genetics*, vol. 10, pp. 252–263, 2009.