Identifying Driver Genes of Breast Cancer by Integrated Analysis of Methylation and Expression Data in Paired Disease-Normal Samples of Patients

Xiaopei Shen, Shan Li, and Zheng Guo

***Abstract*—Among the thousands of gene promoters hyper- or hypomethylated in cancer genomes, only a small portion of them play "driver" roles in tumorigenesis, whereas the others are only "passengers". Here, we develop an approach to identify driver methylation genes of cancer with integrated promoter methylation and gene expression data generated from paired cancer and normal samples for each of a cohort of breast cancer patients, taking the advantage that data of paired samples could provide the relative gene methylation change information from normal to tumor for each individual patient. We applied this approach to analyze a dataset of breast cancer and discovered some novel cancer driver genes. The identified driver genes with methylation alteration may help us to reveal new molecular targets for potential epigenetic therapy.**

***Index Terms*—Methylation, expression, driver gene, breast cancer.**

I. Introduction

Large amount of methylation alterations have been found in cancer genomes [1]. Only a small portion of the thousands of gene promoter with hypermethylation or hypomethylation may play ‘‘driver’’ roles in tumorigenesis [1], [2], whereas many others are only ‘‘passengers’’ [2], [3]. Therefore, it is an important task to identify ‘driver’ genes with methylation alterations for molecular characterization of cancer. Using a knockout experiment, researchers developed an approach to identify a specific type of driver genes for the survival of cancer cells [3]. However, as there are so many high-throughput gene methylation and expression data for cancer, we could integrate them to derive driver information.

Based on the assumption that a driver gene is expected to influence the expression of this gene and a group of downstream genes affecting particular cancer phenotypes, some works have defined driver copy number alterations [4]. Similarly to copy number alteration, methylation alteration at gene promoters does not alter the coding sequences of genes but influences genes' expression. Thus, we try to apply this assumption to derive driver genes from methylation data. Because there are kinds of cancer phenotypes, we could modify the assumption to be that the downstream genes of a driver gene can affect corresponding cancer pathways to disturb cancer phenotypes [5].

Paired disease-normal sample data not only supplies the information of differences between disease and normal samples, but also contains the information of relative methylation change of each gene from normal to disease for each individual patient, and thus it provides a possibility for us to identify the differential methylation at the level of individual patients. Here, based on the above-mention assumption, we put forward an approach to identify cancer driver genes using paired gene methylation and expression data of cancer. We applied this approach to analyze data for breast cancer to derive driver genes.

Manuscript received December 10, 2012; revised February 27, 2013. This work was supported in part by the National Natural Science Foundation of China (grant number 30970668, 81071646, 91029717, 81201702).

Xiaopei Shen and Shan Li are with the Bioinformatics Centre, School of Life Science, University of Electronic Science and Technology of China, Chengdu, 610054, China (e-mail: bioshenxp@gmail.com, shineyong27@yahoo.cn).

Zheng Guo is with the Department of Bioinformatics, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, 350004, China (e-mail: guoz@ems.hrbmu.edu.cn).

II. Materials and Methods

A. DNA Methylation and Gene Expression Data

TABLE I: Samples of Breast Cancer

Batch	Sample size	Platform	
		Methylation	Expression
Batch 61	14:14	HumanMethylation450	Agilent4502A
Batch 72	3:3		
Batch 85	5:5	HumanMethylation27	
Batch 93	16:16		
Batch 96	8:8		
Batch 106	6:6		

The promoter methylation and expression data for breast cancer paired samples were collected from The Cancer Genome Atlas (TCGA) databases (http://tcga-data.nci.nih.gov/tcga) (see Table I). 52 pairs of tumor and paired normal sample extracted from 5 batches were pooled together (Table Ⅰ). The gene promoter methylation data of batch 85, 93, 96, 103 were collected with the Illumina HumanMethylation27 platform, which detected the methylation level of 27,578 CpG loci located within the proximal promoter regions of transcription start sites of 14495 genes. The methylation data of batch 61, 72 were collected with Illumina HumanMethylation450 platform, which detecting the methylation level of over 450,000 CpG loci covering all gene regions, including promoter and gene body. Because the methylation alterations at promoter usually play important roles for genes transcription, we extracted the loci at promoter which were overlapped between HumanMethylation450 and HumanMethylation27 for follow analysis. Using methylated signal intensity (m) and unmethylated signal intensity (u), the methylation level (M) for each CpG locus was calculated by max (m, 0) / (|u|+

|m|+100). We removed unreliable probes whose proportion of detection P-value >0.05 across all the samples is more than 10%. 1,092 CpG loci within promoters of 605 sex chromosome genes were excluded from the analysis to eliminate gender-specific bias.

The expression data of these samples were collected with the normalized data of Agilent4502A platform. Using T-test, genes with adjusted P values less than 0.05 were defined as differentially expressed (DE) genes.

B. Cancer Genes and Protein-Protein Interaction (PPI) Data

2104 cancer genes were extracted from the Cancer Gene F-Census [6], which is a collection of cancer genes from various data sources.

The human PPI data was collected from eight PPI datasets, including MINT, BIND, IntAct, HPRD , MIPS , DIP , KEGG (Kyoto Encyclopedia of Genes and Genomes) [7] and Reactome protein pairs involved in a complex and neighboring reaction . These PPI data were pooled together [8] and compiled an integrated PPI network of 142,583 distinct interactions involving 13,693 human proteins.

C. Discretization of Methylation Profiles for Individual Cancer Samples

Data discretization was used to identify the state of differential methylation for a locus in a sample. First, with the methylation level in 52 normal samples, the standard deviation (SD) of methylation level of each locus was computed. Then, if the methylation change of a locus between paired tumor (MT) and normal sample (MN) is larger than two SD of this locus, the locus was identified as a differential locus (Fig. 1). For instance, we identify the methylation state of locus i in each sample as

$$\begin{cases} if \quad M_{Ti} - M_{Ni} > 2 \times SD_i, & 1 \\ if \quad M_{Ti} - M_{Ni} < -2 \times SD_i, & -1 \\ \quad others, & 0 \end{cases}$$

At last, the methylation profile of cancer samples were translated into a matrix comprising of 1 (hypermethylation), 0 (no differential methylation) and -1 (hypomethylation).

D. Identifying Driver Genes

Based on the assumption that a driver gene is expected to influence the expression of itself and a group of downstream genes to affect particular cancer associated pathways, we identified a locus with methylation alteration as a driver with following three steps.

Step I For each differentially methylated loci, its gene expression should be significantly down- or up-regulated in hypermethylated or hypomethylated cancer samples comparing with the cancer samples with no differential methylation at this locus (T-test, FDR<0.05) (Fig. 1).

Step II We required that the driver methylation alterations have significantly more downstream genes. The downstream genes of a driver were identified as the DE genes between tumor samples with this methylation alteration (hypermethylation or hypomethylation) and the tumor samples with no differential methylation alteration. Random experiments were performed to see whether the number of downstream genes of the driver alteration was significantly more than expected by chance (FDR<0.01). Specifically, we randomly extracted the same number of tumor samples as samples with the methylation alteration and with no differential methylation, and subsequently performing the identification of DE genes for 10,000 times. The P value of the observed number of DE genes was calculated as the percentage of the random numbers exceeding the observed number.

Step III At least one of the cancer associated pathways should be disturbed by downstream genes of a driver methylation alteration (hypergeometric test, FDR<0.05). We selected 16 cancer associated pathways, as listed in Table II, by referring to the pathways annotated in KEGG "pathway in cancer" [7] (Table II).

If a methylation alteration meets the above three requirements, it is defined as a driver methylation alteration. A gene with at least one driver alteration locus was defined as a driver gene.

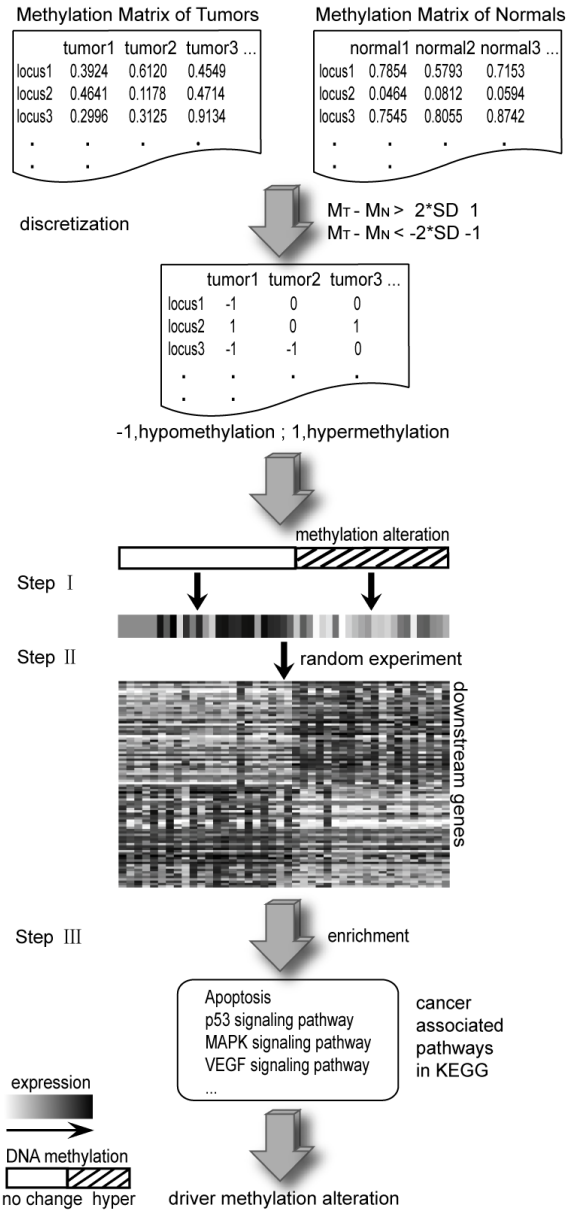


Fig. 1. Schematic overview of the approach to extract driver methylation alterations. The up-regulated and down-regulated genes are labeled with black and white color, separately. The hypermethylation are labeled with black twill.

TABLE II: CANCER ASSOCIATED PATHWAYS

Pathway id	pathway name
hsa04012	ErbB signaling pathway
hsa04150	mTOR signaling pathway
hsa04310	Wnt signaling pathway
hsa04350	TGF-beta signaling pathway
hsa04370	VEGF signaling pathway
hsa04630	Jak-STAT signaling pathway
hsa04110	Cell cycle
hsa04115	p53 signaling pathway
hsa04210	Apoptosis
hsa04510	Focal adhesion
hsa04520	Adherens junction
hsa03320	PPAR signaling pathway
hsa04512	ECM-receptor interaction
hsa04060	Cytokine-cytokine receptor interaction
hsa04151	PI3K-Akt signaling pathway
hsa04010	MAPK signaling pathway

III. RESULT

A. Identification of Driver Genes for Breast Cancer

After data discretization for the methylation profiles of 52 tumor samples (see *Methods*), we restricted our following analysis to 11048 methylation altered loci which were hypermethylated or hypomethylated in at least 10% of all cancer samples. Firstly, using T-test with FDR<0.05, we identified 143 loci hypermethylated or hypomethylated within the promoters of 131 genes that were significantly down-regulated or up-regulated in cancer samples. Secondly, from these 143 loci, we found 60 loci of 57 genes which influenced the expression change of significantly more downstream genes than expected by random chance according to the random experiments described in the *Methods*. Finally, from these 60 loci, we identified 33 loci of 32 genes whose downstream genes were significantly enriched in at least one of the cancer-associated pathways defined in "pathway in cancer" (hypergeometric test, FDR<0.05). Each driver gene was outputted with corresponding disturbed pathways (Table III).

TABLE III: DRIVER GENES AND THE PATHWAYS THEY DISTURBED

driver gene	pathway name
SPDEF	ErbB signaling pathway
FAM128B, SPDEF	mTOR signaling pathway
MFAP4, NME5, HOXB8, **ERBB3**, **MGAT1**, SLC2A5, TNFRSF1B, ARHGAP30, SP140, HMGCS2, CD37, TMEM149, **PTPRCAP**, **TLR9**, F7, FAM128B, MB, **CCR1**, SPDEF, **ARHGAP25**	Jak-STAT signaling pathway
SLC9A11, **RBP1**, SHROOM1, CEACAM19, ZNF454, IL1R2, CAPN9	Cell cycle
ZNF454, IL1R2	p53 signaling pathway
MB,**CCR1**, SPDEF	Apoptosis
MFAP4, NME5, HOXB8, **ERBB3**, **MGAT1**, SLC2A5, TNFRSF1B, ARHGAP30, SP140, HMGCS2, CD37, TMEM149, **PTPRCAP**, **TLR9**, F7, EOMES, SH2D3C, **CCND1**, **CD6**, CAPN9, FAM128B, MB, **CCR1**, SPDEF	Cytokine-cytokine receptor interaction

*Gene names in bold were known cancer genes from F-census

B. Validation of the Identified Driver Genes

Evidences supported that these driver genes are likely to play driver roles in tumorigenesis. Firstly, 9 (28.1%) of the identified 32 driver genes were known cancer genes collected in the F-census database [6], which was significantly more than what expected by random chance (hypergeometric test, P=1.07E-04) (Table II). For an example, tumor suppressor gene RBP1 is a regulator of breast epithelial retinoic acid recepter activity, cell differentiation, growth arrest and cell cycle progression [9], [10], and it was identified as a driver genes with promoter hypermethylation disturbing "cell cycle". As another example, CD6 was identified as a driver genes with promoter hypomethylation, in accordance with previous report that increased expression of CD6 suppresses longer term events such as cytokine secretion and T-cell proliferation [11], which could promoter the initiation of cancer. Secondly, in addition to the known cancer genes collected in the F-census database, some driver genes have been suggested to be cancer genes in previous studies. For instance, IL1R2 has been identified as a driver gene with promoter hypomethylation, in accordance with a previous report that this gene is a possible cancer gene [12].

Then, after removing the 9 known cancer genes from the 32 identified driver genes, we found that the 9 of the remaining 23 driver genes had at least one PPIs connected with known cancer genes in F-census. This result implied that some of the newly predicted driver genes might work closely with the known cancer genes and they might perform similar functions as their neighboring cancer genes in tumorigenesis. For instance, it has been reported that cancer genes IL1A, IL1B could activate NF-kB signal pathway, and disturb cell cycle mediators [13]. In our analysis, their neighbouring gene IL1R2 was identified as a hypomethylated driver gene with its downstream genes disturbing cell cycle and TP53 pathways.

IV. DISCUSSION

Methylation alteration in cancer genome is a widely molecular change, but how to extract the driver methylation alterations of cancer is still a problem. The identification of driver genes with methylation alterations and pathways they disturbed is a fundamental step towards mechanistic characterization of cancer and may provide potential targets of epigenetic therapy considering the reversibility of methylation [14]. In this study, we proposed a computational approach to identify driver genes by taking into account not only the association between promoter methylation and gene expression but also the association between a candidate driver and its downstream genes. Also, the pathways represented by the downstream genes can help us to gain insight into how a driver methylation alteration contributes to the malignant phenotype by altering the cellular pathways.

Data of paired sample usually supply more reliable information. The individual analysis of relative changes for paired samples could catch the actual alterations of each locus for each patient. It also supplies a new way to dip further information from data of paired samples.

Additionally, batch effects, which could be introduced by using samples from different experimental batches, may produce systematic non-biological differences between different groups of samples [15]. In our work, the normal and tumor sample in a pair were both come from the same batch, so relative changes of methylation level were not influenced by batch effects. However, the SD of methylation level of

each locus was possibly enlarged by batch effects, and it might bring a too strict threshold for discretization, which influence the power of discovering driver genes.

The limitation of our method is that currently there is no widely accepted definition of cancer pathways. The cancer pathways we selected were all come from "pathway in cancer" in KEGG [7]. As the improvement of definition for cancer pathways, the performance of our approach would be improved. Finally, except for methylation alteration, mutation and copy number change can also influence the expression of driver genes. Thus, we will try to integrate these types of molecular alterations and improve the approach to identify driver genes of cancer in our future work.

REFERENCES

[1] X. Shen *et al.*, "Distinct functional patterns of gene promoter hypomethylation and hypermethylation in cancer genomes," *PLoS One,* vol. 7, no. 9, pp. e44822, 2012.

[2] S. Kalari and G. P. Pfeifer, "Identification of driver and passenger DNA methylation in cancer by epigenomic analysis," *Adv Genet,* vol. 70, pp. 277-308, 2010.

[3] D. D. De Carvalho *et al.*, "DNA methylation screening identifies driver epigenetic events of cancer cell survival," *Cancer Cell,* vol. 21, no. 5, pp. 655-67, May 15, 2012.

[4] U. D. Akavia *et al.*, "An integrated approach to uncover drivers of cancer," *Cell,* vol. 143, no. 6, pp. 1005-17, Dec. 10, 2010.

[5] S. Efroni, C. F. Schaefer, and K. H. Buetow, "Identification of key processes underlying cancer phenotypes using biologic pathway analysis," *PLoS One,* vol. 2, no. 5, pp. e425, 2007.

[6] X. Gong *et al.*, "Extracting consistent knowledge from highly inconsistent cancer gene data sources," *BMC Bioinformatics,* vol. 11, pp. 76, 2010.

[7] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res,* vol. 28, no. 1, pp. 27-30, Jan. 1, 2000.

[8] K. Lage *et al.*, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nat Biotechnol,* vol. 25, no. 3, pp. 309-16, March 2007.

[9] A. Lai, R. C. Marcellus, H. B. Corbeil, and P. E. Branton, "RBP1 induces growth arrest by repression of E2F-dependent transcription," *Oncogene,* vol. 18, no. 12, pp. 2091-100, March 25, 1999.

[10] E. F. Farias *et al.*, "Cellular retinol-binding protein I, a regulator of breast epithelial retinoic acid receptor activity, cell differentiation, and tumorigenicity," *J Natl Cancer Inst,* vol. 97, no. 1, pp. 21-9, Jan. 5, 2005.

[11] M. I. Oliveira *et al.*, "CD6 attenuates early and late signaling events, setting thresholds for T-cell activation," *Eur J Immunol,* vol. 42, no. 1, pp. 195-205, Jan. 2012.

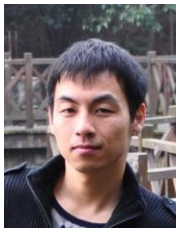
[12] F. Ruckert *et al.*, "Examination of apoptosis signaling in pancreatic cancer by computational signal transduction analysis," *PLoS One,* vol. 5, no. 8, pp. e12243, 2010.

[13] G. Multhoff, M. Molls, and J. Radons, "Chronic inflammation in cancer development," *Front Immunol,* vol. 2, pp. 98, 2011.

[14] J. P. Issa, "DNA methylation as a therapeutic target in cancer," *Clin Cancer Res,* vol. 13, no. 6, pp. 1634-7, March 15, 2007.

[15] J. T. Leek *et al.*, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nat Rev Genet,* vol. 11, no. 10, pp. 733-9, Oct, 2010.

Xiaopei Shen was born on October 30, 1983, in Anhui, China. He received the B.S. degree in computer science from Centre South University and M.S. degree in Biophysics from University of Electronic Science and Technology of China (UESTC). He is currently a Ph.D. candidate in studying in bioinformatics in UESTC. He is interested in gene mutation, expression and methylation analysis in cancer genome.



Shan Li was born on February 7, 1989, in Hebei Province. She graduated from Harbin Medical University and gained a bachelor's degree in Bioinformatics. Now she is a postgraduate student in Biophysics of University of Electronic Science and Technology of China (UESTC). She is focused on expression and methylation analysis in cancer genome.

Zheng Guo was born on October 9, 1963. He obtained the PHD degree in computer science from Harbin Institute of Technology and is currently a Professor of Bioinformatics at Fujian Medical University. He has undertaken multiple National Natural Science Foundation projects. His research is focused on analysing microarray data and complex diseases at the functional module level.