Inferring genome-wide interplay landscape between DNA methylation and transcriptional regulation

Binhua Tang[1, 2]* and Xin Wang[2]
[1]The Internet of Things School, Hohai University, Jiangsu, China
[2]Harvard Medical School, Harvard University, Boston, MA

Abstract: DNA methylation and transcriptional regulation play important roles in cancer cell development and differentiation processes. Based on the currently available cell line profiling information from the ENCODE Consortium, we propose a Bayesian inference model to infer and construct genome-wide interaction landscape between DNA methylation and transcriptional regulation, which sheds light on the underlying complex functional mechanisms important within the human cancer and disease context. For the first time, we select all the currently available cell lines (>=20) and transcription factors (>=80) profiling information from the ENCODE Consortium portal. Through the integration of those genome-wide profiling sources, our genome-wide analysis detects multiple functional loci of interest, and indicates that DNA methylation is cell- and region-specific, due to the interplay mechanisms with transcription regulatory activities. We validate our analysis results with the corresponding RNA-sequencing technique for those detected genomic loci. Our results provide novel and meaningful insights for the interplay mechanisms of transcriptional regulation and gene expression for the human cancer and disease studies.

Keywords: Interplay mechanism; DNA methylation; transcriptional regulation; tumor; cell line.

INTRODUCTION

Aberrant DNA methylation is frequently documented as its complicated association to cancers and diseases. While till now there still lacks of systematic investigation of the cell-specific DNA methylation, transcriptional regulation mechanism, genome-wide detection of risk loci and clinical association in human cancers (Reik *et al*., 2001; Tang *et al*., 2013a; Smith and Meissner, 2013; Tang *et al*., 2013b; Wang *et al*., 2014).

Through systematic integration of DNA methylation and transcriptional regulation binding information, we construct a global interaction landscape for the both functional mechanisms on gene expression process across the 20 diverse cell and tissue types and 82 transcription factors, currently available in the ENCODE Consortium portal (Consortium, 2011; Consortium, 2012).

Genome-wide DNA methylation status, transcriptional regulation binding information and gene expression profiles across multiple tumorous and normal cell lines are analyzed using the proposed Bayesian multilevel model, thus we can further identify and quantify the impacts on downstream mechanisms on gene expression activities by both transcriptional regulation and DNA methylation.

Furthermore with benchmark evaluation of two main breast cancer cell lines, MCF-7 and T-47D (Lin *et al*., 2004; Zullo *et al*., 2012; Tang *et al*., 2012), we have identified cell-specific differential methylation and transcription binding loci, and we find most of those loci residing in the vicinity of such key transcription factors as CTCF and GATA3 (Handoko *et al*., 2011; Yagi *et al*., 2011).

The newly identified risk loci with genomic information are then annotated; then with the RNA-sequencing platform data source, we further analyze and validate the differentially-expressed genes within those identified genomic loci.

MATERIAL AND METHODS

To infer the interplay landscape between DNA methylation and transcriptional regulation, we select the corresponding 20cell line profiling data sources currently available from the ENCODE Consortium portal, ***i.e.*** DNA methylation profiling data of two platforms, Illumina Infinium 450K and the Reduced Representation Bisulfite Sequencing (RRBS), the transcription factor binding information (ChIP-sequencing, ChIP-seq) (Park, 2009; Valouev *et al*., 2008) and differential gene expression data sources (Whole Transcriptome Shotgun Sequencing, RNA-sequencing, RNA-seq) (Haas and Zody, 2010; Wang *et al*., 2009).

Since the two platforms are profiled with the quite different techniques, we examine their genome-wide profile distribution property in fig. 1. The correlation coefficient denotes the similarity for two platform techniques on the same investigated object.

**Corresponding author:* e-mail: bh.tang@outlook.com

Then for quantifying the differential methylation status residing in diverse genome-wide loci. We dissect the genome into six non-overlapping regions based on the genomic promoter, CpG island (CGI) and CpG island shore (CGIS), etc. Then we quantify the methylation profile status interacting with the transcriptional regulation activities within those predefined genomic regions. Fig. 2 illustrates the methylation distribution status for the dissected genomic regions.

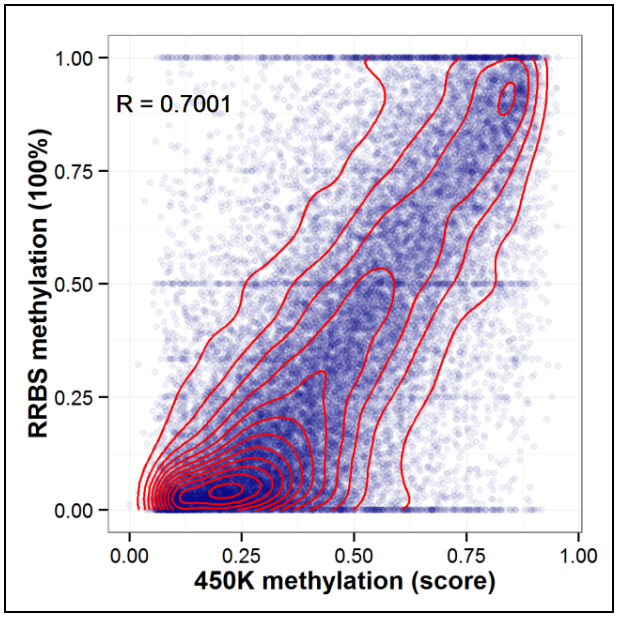

Fig. 1: Genome-wide DNA methylation status comparison by the two platforms (Illumina Infinium 450K and RRBS) for the T-47D cell line. Pearson correlation coefficient measuring the profile similarity is given on the top left. Contour line represents the methylation level intensity distribution.

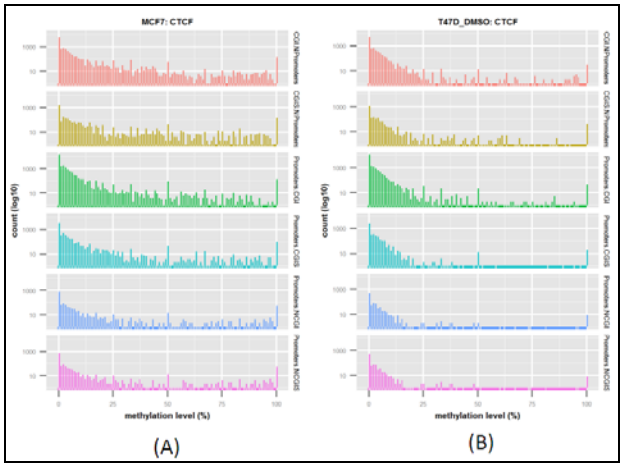

Fig. 2: Methylation status within six dissected genomic regions of interest for MCF-7 (A) and T-47D (B) cell lines. The two cell lines are differentially-methylated at the high methylation level, esp. for the top first three regions.

Furthermore we propose a Bayesian inference model for quantifying the interplay mechanisms on gene expression status by the interplay activities between transcriptional regulation and DNA methylation, the Bayesian model is depicted as below,

$$E_k[G] = \sum_{i=1}^{R} M_{ki}[G] + \sum_{j=1}^{R} T_{kj}[G] + \Phi \qquad (1)$$

$k \in N,\ R \in N,\ \Phi \sim ¥(0, \sigma^2)$

where E denotes transcript expression status, G is a gene set matrix, M for methylation level and T for transcription regulatory activity for the gene set G, k is the gene set size, and R is the quantity of dissected regions.

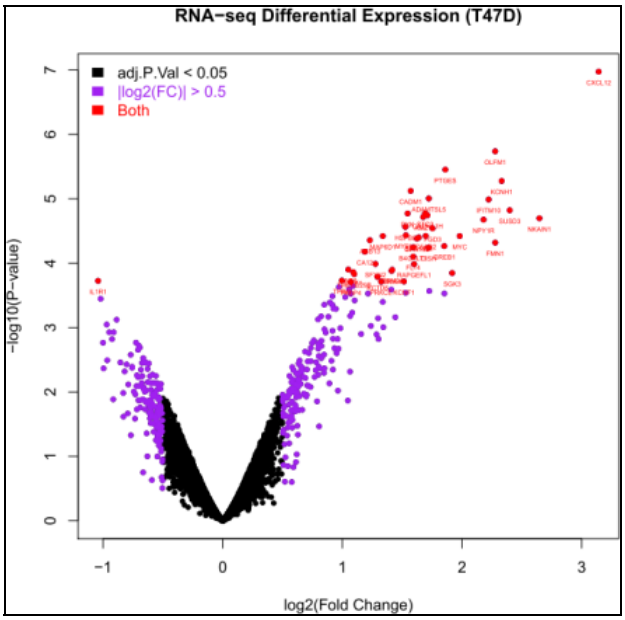

Fig. 3: Differential transcript expression analysis for the genes within the functional genomic regions from the T-47D cell line using the RNA-seq technique platform. The black dots for the genes with adjusted P-value <0.05 only, the purple dots for those genes with the absolute log2 fold change >0.5 only and the red dots for those genes satisfying the both criteria, respectively.

ANALYSIS RESULTS

Transcriptional regulation is frequently documented as its association with DNA methylation. One of the main reasons could be that TF-DNA binding is often sequence-specific. Therefore, DNA methylation at those CTCF binding sites disrupts the binding of CTCF to DNA, hence dysregulates the expression of CTCF-regulated genes. The conclusion has also be confirmed in the recent work (Wang *et al*., 2012).

Based on the analysis, we find the genomic responses triggered by the interplay status between DNA methylation and transcriptional regulation differ for diverse regions, according to the analysis results on those

cell lines. A good case in point is transcription factor CTCF for the MCF-7 and T-47D cell lines.

Furthermore, the interplay between CTCF occupancy and DNA methylation is known to be cell-specific. For MCF-7 cell line, the methylation level is highly interplaying with the transcriptional regulation in CGI-only and CGIS-only regions, rather than the promoter-only regions, compared with T-47D cell line.

Using the Whole Transcriptome Shotgun Sequencing (WTSS) technique, *i.e*. RNA-seq platform data, we further analyze the differential transcript expression status for the genes located within the dissected regions of interest.

Fig. 3 illustrates the differential transcript expression status for those genes of interest, where the adjusted *P*-value threshold (adj.P.Val) is 0.05 and the absolute log2 fold change (|log2(FC)|) threshold is 0.5. The red dots denote those genes satisfying the both thresholds, *i.e.* adj.P.Val <= 0.05 and |log2(FC)| >=0.5; the black dots for the genes with adj.P.Val < 0.05 only, and the purple dots for those genes with |log2(FC)| >0.5 only, respectively.

Based on the above analysis, we quantified those genes of interest in diverse genomic regions with regard to the DNA methylation status, especially for the differential methylation levels in the CpG island and promoter regions.

CONCLUSIONS AND PERSPECTIVE

A Bayesian inference model is proposed to quantify the interplay mechanisms on gene expression by both transcriptional regulation and DNA methylation. Through the proposed method together with the integrated multiple genomic information from the ENCODE Consortium portal, we have detected multiple genomic loci of interest, and further quantified the differential gene expression status, which directly impacted by the interplay mechanisms between DNA methylation and transcription activity.

The analysis results were further validated by the NGS-based platform, Whole Transcriptome Shotgun Sequencing (WTSS), *i.e*. RNA-seq technique for quantifying transcript expression profile. Multiple genes located within the hotspots detected from the DNA methylation and transcript expression profiles were further quantified differentially. Our future work will emphasize on the discovery of more genomic and clinical information with a quantitative measure. Those information paves the way for the current translational study with application to cancer and clinical research. Another very promising way is to design and integrate more genetics and epigenetics information using a systematic and efficient approach, which will provide predictable insights into the underlying mechanism complexity.

ACKNOWLEDGMENT

Special thanks to the editors and reviewers for their valuable comments. The study is supported by the research funding under grant No. 2012B03914 (BZX12B101-02) and XZX/10B007-02 from Hohai University, and the postdoctoral research funding from Harvard Medical School.

REFERENCES

Consortium TEP (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS. Biol.,* **9**: e1001046.

Consortium TEP (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**: 57-74.

Haas BJ and Zody MC (2010) Advancing RNA-seq analysis. *Nat. Biotech.,* **28**: 421-423.

Handoko L, Xu H and Li G *et al* (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet,* **43**: 630-638.

Lin C-Y, Strom A and Vega V *et al* (2004). Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells. *Genome Biology,* **5**: R66.

Park PJ (2009). ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet,* **10**: 669-680.

Reik W, Dean W and Walter J (2001). Epigenetic reprogramming in mammalian development. *Science,* **293**: 1089-1093.

Smith ZD and Meissner A (2013). DNA methylation: Roles in mammalian development. *Nat. Rev. Genet,* **14**: 204-220.

Tang B, Gu F and Jin VX (2013a). Inference of gene regulatory networks in breast and ovarian cancer by integrating different genomic data. Statistical Diagnostics for Cancer. Wiley-VCH Verlag GmbH & Co. KGaA, pp.153-171.

Tang B, Hsu HK and Hsu PY *et al* (2012). Hierarchical modularity in ERα transcriptional network is associated with distinct functions and implicates clinical outcomes. *NPG Scientific Reports,* **2**: 875.

Tang B, Hsu PY and Huang THM *et al* (2013b). Cancer Omics: From regulatory networks to clinical outcomes. *Cancer Letters.* **340**: 277-283.

Valouev A, Johnson DS and Sundquist A *et al* (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Meth.,* **5**: 829-834.

Wang H, Maurano MT and Qu H *et al* (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research,* **22**: 1680-1688.

Wang Z, Curry E and Montana G (2014). Network-guided regression for detecting associations between DNA methylation and gene expression. *Bioinformatics,* **30**: 2693-2701.

Wang Z, Gerstein M and Snyder M (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet,* **10**: 57-63.

Yagi R, Zhu J and Paul WE (2011). An updated view on transcription factor GATA3-mediated regulation of Th1 and Th2 cell differentiation. *International Immunology,* **23**: 415-420.

Zullo Joseph M, Demarco Ignacio A and Piqué-Regi R *et al* (2012) DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell,* **149**: 1474-1487.