Exploring Regulatory Elements in Low-methylated Regions for Gene Expression Prediction

Hong Hu* and Yang Dai*

Abstract— **Recent studies on methylomes have indicated that low-methylated regions (LMRs) are related to potential active distal regulatory regions. To further investigate the potential relation between LMRs and gene expression regulation, we propose a penalized logistic regression model to predict gene expression directional change based on computationally analyzed transcription factor binding sites in LMRs that are distinctive between two cell types. We evaluated this approach using the whole genome bisulphite sequencing and RNA-seq data of two cell types: adipose-derived stem cells and iPSCs of adipose-derived stem cells. For Differentially Expressed (DE) genes with LMRs in their intergenic and/or genebody regions, our model obtained a** 10**-fold cross-validated AUC value of** 0.88 **for prediction of expression directional change. For DE genes with only LMRs in intergenic regions the corresponding AUC value is** 0.84**.**

I. INTRODUCTION

The application of the whole genome bisulphite sequencing (WGBS) technology has led to the identification of low-methylated regions (LMRs) in a methylome of a particular cell type [1], [2]. LMRs reside in the CpG-poor regions with an average methylation level of 30% and occur mostly distal to promoters. These regions are strongly enriched for chromatin features such as DNase I hypersensitivity, high H3K4 monomethylation (H3K4mel) signal relative to H3K4 trimethylation (H3K4me3) and the presence of p300 histone acetyltransferase [1], which are indicative feature of enhancers [3]. These facts imply that some of the LMRs are active distal regulatory regions. In addition, the binding of transcription factors (TFs) to DNA was found necessary to create LMRs [1], [4]. These findings suggest that this unique DNA methylation feature could be used as a means to detect the distal enhancer regions from which TFBSs could be identified to infer the regulatory mechanism. The detected LMRs in one sample could be compared with that in a different sample to identify differential LMRs (dLMRs).

Common approaches for characterizing methylation changes between two samples were proposed before [6], [7], [8], where a sliding window is used to identify differentially methylated regions (DMRs). It has been shown that there is a weak anticorrelation between gene expression level and the level of methylation in the DMR near the promoter of a gene. A recent work has revealed a diverse set of patterns of methylation level that are strongly associated gene expression [9] using a more sophisticated analysis of methylation change in windows of various size around the transcription starting sites (TSSs) of genes. However, non-promotor DMRs seem to cover more genomic regions [5].

Our current work is based on the hypothesis that the dLMRs between two samples are associated to gene expression changes between the two samples. If the dLMRs indeed include active distal regulatory regions, it would be meaningful to explore regulatory elements and to evaluate the usefulness of this information in prediction of the gene expression. Towards this goal, we propose to use a penalized logistic regression model to address the following question,

- Can we predict gene expression change using the computationally analyzed TFBSs in the dMLRs?

II. MATERIALS AND METHOD

A. Detection of cell-type specific LMRs

The first step in WGBS data analysis is to align the bisulfite converted reads and quantify the methylation level for individual Cytosines(C) in the genome. This analysis can be accomplished by using packages such as *Bismark* [11]. The methylation percentage for each C is defined as the ratio of the number of alignments with C (methylated), over the number of alignments with either C (methylated) and G (unmethylated). This quantification provides the methylome of a sample. The LMRs are detected from the analysis of methylome using the package *MethylSeekR* [2]. The identified LMRs are associated to the nearest genes using *ChIPpeakAnno* [12]. As we stated before, LMRs are active regulatory regions which may have a role in regulation of expression of the genes.

The next step is to identify dLMRs between two sample groups. The dLMRs are defined as LMRs in one sample type that do not overlap with the LMRs in the other sample type. In addition, if some dLMRs are associated with a single gene and the corresponding LMRs are from different samples, then those dLMRs and genes are excluded from further analysis. By completing this procedure, we identify a set of genes that are associated with dLMRs detected from each sample type, but not both.

B. Identification of TFBSs in dLMRs

To assess the potential regulatory effect of dLMRs on gene regulation, we exam the relationship between the predicted TSBSs and the gene expression change between two samples. First, the Position Weight Matrices (PWMs) are used to predict the TFBSs in each dLMR. The statistically significant TFBSs in individual dLMRs are determined based on empirical distributions of similarity scores obtained from randomly generated sequences. More specifically, the frequencies of

* Department of Bioengineering (MC563), University of University of Illinois at Chicago, 835 S. Wolcott St, W100 CSN Chicago, IL 60612, USA {hhu4,yangdai} at uic.edu

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

DNA nucleotides are calculated from the identified dLMRs and used to generate random background sequences with a 0-order Markov model. The PWMs are used to scan the random sequences to obtain empirical distributions of similarity scores for individual PWMs. The significant TFBSs for each PWM are identified with a prescribed threshold for p-values adjusted by the Bonferroni procedure. If there are several significant TFBSs for a single PWM in a dLMR, then the mean value of similarity scores of all these TFBSs is considered as the similarity score for the PWM in the dLMR.

In this study we use the 123 non-redundant *Homo sapiens* PWMs released in 2014 JASPAR database [13]. The choice of non-redundant PWMs eliminates the complications where multiple PWMs may represent a TF, or a PWM represents multiple TFs. Therefore, in the remainder of the paper, PWMs or TFs are used interchangeably. The prediction of TFBSs is carried out by using the Bioconductor package *PWMEnrich* [14]. We generate $1,000$ random sequences of $1,000$bp long. The threshold for the adjusted p-values is 0.05.

C. Identification of DE Genes

Since the number of replicates is usually small in RNA-seq data, the Bioconductor package *NOISeq* is chosen for the identification of DE genes between two cell types [15]. Briefly, short read counts for genes are converted into Reads Per Kilobases per Million (RPKM), and genes with RPKM $<$ 1 are filtered from further analysis. The *NOISeq* empirically models the noise in count data and is reasonably robust against the choice of standard deviation (SD). The DE genes are defined as those with probability ($\geq$ 0.8) of being differentially expressed given expression log ratio (M) and absolute value of difference (D) [15].

D. Penalized Logistic Regression Model for Prediction of Direction Change of Gene Express

The penalized logistical regression model is used to evaluate the association between the TFBSs identified in dLMRs and the direction change of gene expression. We take a subset of DE genes that are associated with at least one dLMR. The number of genes is denoted as N_g. Each gene is assigned a similarity score for each PWM based on the scanning results described above. If a gene has multiple dLMRs, the maximum of the scores from all the dLMRs is used as the similarity score for that gene. If a PWM does not have significant TFBS in a dLMR, the similarity score is 0. This information can be organized into a matrix X of size $N_g \times N_{TF}$, where N_{TF} denotes the number of PWMs involved. In addition, let $\mathbf{Y}$ be the vector of the labels for genes, i.e., $Y_i = 1$ if gene i is up-regulated and $Y_i = 0$ if down-regulated with the reference sample types. Now we can model the gene expression labels by using the logistic regression model as follows,

$$\log \frac{Pr(Y_i = 1|X_i)}{Pr(Y_i = 0|X_i)} = X_i\beta + \beta_0, \quad i = 1, ..., N_g \qquad (1)$$

where X_i is the i^{th} row of matrix X; β is a column vector of coefficients. The logistic regression coefficients are typically estimated by the maximum likelihood method [16].

It is likely that not all PWMs are informative in the model. Therefore, the penalized logistic regression model is more suitable. This model amounts to minimize the following function:

$$L(\beta_0, \beta, \lambda) = -\ell(\beta_0, \beta) + \frac{\lambda}{2}\|\beta\|^2 \qquad (2)$$

where ℓ indicates the binomial log-likelihood, and λ is a positive constant that needs to be determined. We solve this model by using the R package *glmnet* [17] through a 10-fold cross-validation procedure.

III. Results

Our analysis used two sets of WGBS data that were previously reported for the two cell types: adipose-derived stem cells (*ads*) and iPSC of the adipose-derived stem cells (*ads-ipsc*) [18]. The aligned WGBS data and RNA-seq data for both cell types were obtained from *http://neomorph.salk.edu/ips_methylomes/data.html*. The RNA-seq data includes two replicates for each cell type.

LMRs and dLMRs

By using *MethylSeekR*, $47,618$ and $13,424$ LMRs were identified for *ads* and *ads-ipsc* respectively. We associated each LMR to its closest gene annotated by *RefSeq* IDs. The distributions of LMRs relative to the TSSs of their associated genes are shown in Figure 1. For both cell types, it can be seen that large proportions of LMRs are located in regions more than 10kb away from the TSSs.

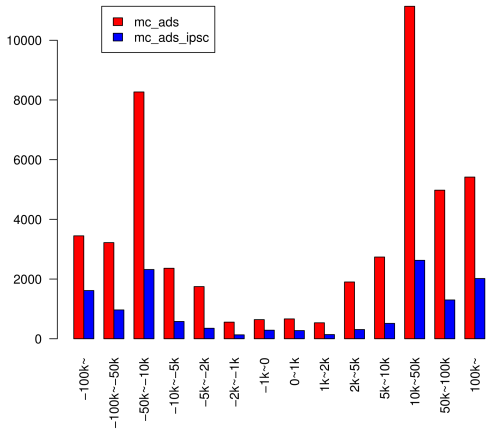

Fig. 1. The distribution of LMRs to their associated genes.

We divided the LMRs into three categories based on their distances to the TSSs of the associated genes. (1) Promoter: LMRs that overlap the promoter regions ($1,000$ bp upstream and 500 bp downstream of TSSs). (2) Genebody: LMRs that overlap the genebody regions. (3) Intergeneic: LMRs that overlap the intergenic regions, i.e., LMRs not overlap promoter or genebody regions. Table I provides the LMR distributions for the two cell types. It is noticable that most of the LMRs fall in genebody and intergenic Regions.

TABLE I

THE NUMBER OF IDENTIFIED LMRS AND THEIR GENOMIC FEATURES.

Cell type	Total	Promoter	Genebody	Intergenic
ads	47,618	1,374	15,374	30,870
ads-ipsc	13,434	547	3,632	9,255

Figure 2 shows the boxplots for the sizes and CpG contents of LMRs. The results confirmed that the LMRs are CpG poor, and a large proportion of them locates in intergenic as well as genebody regions.

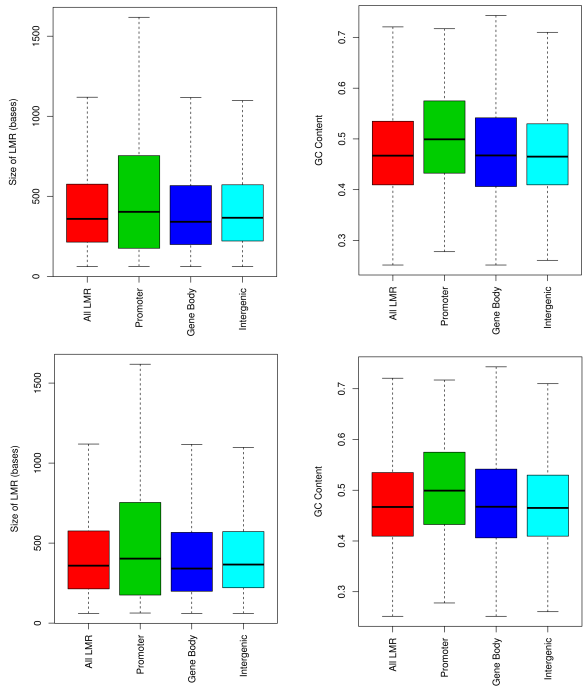

Fig. 2. The boxplots of LMRs size and GC content for *ads* (top) and *ads-ipsc* (bottom).

Following the procedure described previously, we identified the dLMRs between the two cell types. 26,723 dLMRs were observed in *ads* only and associated with 9,502 unique RefSeq IDs; 3,871 dLMRs were observed in *ads-ipsc* only and associated with 2,483 RefSeq IDs. Further analysis revealed that only a small number of genes having dLMRs in their promoter regions and most of genes are associated with dLMRs located in genebody and intergenic regions. Table II shows the details. Note that one gene may associated with multiple dLMRs in different genomic regions.

TABLE II

NUMBER OF REFSEQ IDS THAT ARE ASSOCIATED WITH DLMRS IN DIFFERENT GENOMIC REGIONS

Cell type	Total	Promoter	Genebody	Intergenic
ads	9,502	896	3,837	4,769
ads-ipsc	2,483	197	721	1,565

DE genes

We identified 2,298 DE genes between *ads* and *ads-ipsc* from the RNA-seq data set using *NOISeq* from *Bioconductor*. Out of which 480 genes had at least one dLMR. Since dLMRs in distal and genebody regions are of our primary interest, we further removed genes with dLMRs located in promoter regions. Two DE genes sets were prepared in order to evaluate the association levels of dLMRs from different regions to gene expression change. Set1 includes the DE genes with only intergentic dLMRs. This set consists of 250 up-regulated and 18 down-regulated genes in *ads* compared to *ads-ipsc*. Set2 is comprised of DE genes with either intergenic and/or genebody dLMRs. This set includes 410 up-regulated and 28 down-regulated genes for *ads* comparing to *ads-ipsc*.

We trained the penalized logistic regression model with a 10-fold cross-validation procedure to identify the best value for λ for each dataset. The Area Under the *ROC* Curve (AUC) was used as the criterion for performance evaluation. Figure 3 shows the cross-validated AUC values of the model at different values of λ for each dataset. As observed the predictions at the best parameter values are quite promising; giving a AUC value of 0.84 for Set1 and 0.88 for Set2.

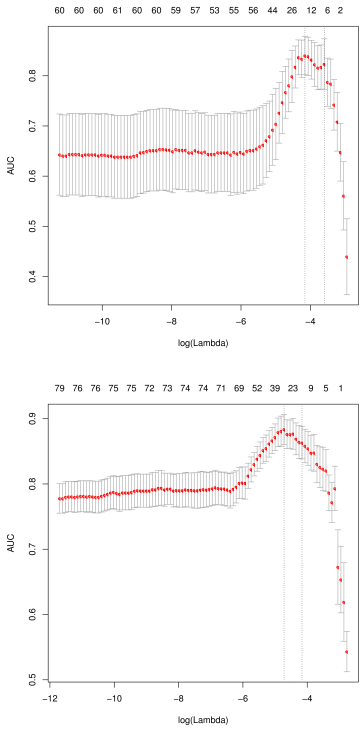

Fig. 3. The cross-validated AUC values vs. $\log(\lambda)$ values. Top: DE genes with Intergenic dLMRs. Bottom: DE genes with Intergenic and Genebody dLMRs.

Finally, we used the following procedure to identify a stable set of non-zero β coefficients. The best λ value for each data set was determined based on the 10-fold cross-validation procedure. Using the best λ value, the model was trained using each dataset for 100 times. The average of $\bar{\beta}$ vectors learned from the 100 runs was calculated. From the

averaged β vectors obtained for Set1 and Set2, we identified 26 PWMs corresponding to non-zero coefficients for Set1; and 76 PWMs with non-zero β coefficients for Set2. We further removed PWMs with coefficients of extremely small asolution values ($< 10^{-5}$), resulting in 18 PWMs for the Set1 and 72 PWMs for Set2.

For the selected TFs, we further checked if the predicted regulations are supported by literature. For DE genes in Set1, we successfully detected the binding of TF *FOSL1* on gene *JUN* [19]. *FOSL1* encodes leucine zipper proteins that can dimerise with proteins of the JUN family, thereby forming the TF complex *AP-1*. Another example is the binding of TF *TP63* on *N4BP2* gene [20] in Set2. *TP63* is a member of *p53* family and has been related with apoptosis [21].

IV. Discussion and Conclusion

We have developed a penalized logistic regression model to evaluate the regulatory effects of TFBSs in low-methylated regions identified from the analysis of WGBS data. Our preliminary results demonstrate the promising performance of our model, suggesting the similarity scores of TF binding sites in LMRs in intergenic and/or genebody are indeed predictive for gene expression directional changes. This computational analysis provided further supportive evidence that TF binding sites in LMRs in distal regions or genebody may be functional, implying potential distal regulatory mechanisms of the LMRs.

The identification of distal regulatory regions is one of the major challenges in gene regulation study. The ChIP-seq based experimental approach to the detection of global TF binding sites is useful, however, it is limited by antibody availability. Other epigenetic features, such as chromatin structure and histone modification marks have been utilized for the prediction of active regulatory regions including enhancers [22]. These studies have revealed that the histone modification patterns are far from sufficient for such a task [23]. The integration of data from chromatin structure, histone modification and DNA methylation for a comprehensive enhanser study would be a future direction.

References

[1] Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, and Schubeler D. "DNA-binding factors shape the mouse methylome at distal regulatory regions," Nature, 2011, 480, (7378):90-495.

[2] Burger L, Gaidatzis D, Schubeler D, and Stadler MB."Identification of active regulatory regions from DNA methylation data", Nucleic Acids Research, 2013, 41(16):e155.

[3] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, and Ren B. "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome", Nat Genet, 2007, 39 (3):311-318.

[4] Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, and Schubeler D. "Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at Regulatory Regions", PLoS Genet, 2013, 9 (12):e1003994.

[5] Berman B, Weisenberger D, Aman J, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange C, van Dijk C, Tollenaar R, Van Den Berg D, and Laird P. "Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains", Nat Genet, 2012, 44:40 - 46.

[6] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q.-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, and Ecker JR. "Human DNA methylomes at base resolution show widespread epigenomic differences", Nature, 2009, 462 (7271): 315-322.

[7] Hansen K, Timp W, Bravo H, Sabunciyan S, Langmead B, McDonald O, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry R, and Feinberg A. "Increased methylation variation in epigenetic domains across cancer types", Nat Genet, 2011, 43:768 - 775.

[8] Hansen K, Langmead B, and Irizarry R. "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions", Genome Biology, 2012, 13 (10):R83.

[9] VanderKraats ND, Hiken JF, Decker KF, and Edwards JR. "Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes", Nucleic Acids Research, 2013, 41 (14):6816-6827.

[10] Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CPE, van Dijk CM, Tollenaar RAEM, Van Den Berg D, and Laird PW. "Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains", Nat Genet, 2012, 44 (1):40-46.

[11] Krueger F, and Andrews SR. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications", Bioinformatics, 2011, 27 (11): 1571-1572.

[12] Zhu L, Gazin C, Lawson N, Pages H, Lin S, Lapointe D, and Green M. "ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data", BMC Bioinformatics, 2010, 11 (1):237.

[13] Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C.-y, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, and Wasserman WW. "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles", Nucleic Acids Research, 2014, 42(Database issue):D142-7

[14] Stojnic R, Diez D. "PWMEnrich: PWM enrichment analysis", R package version 2.6.2, 2013.

[15] Tarazona S, Garca-Alcalde F, Dopazo J, Ferrer A, and Conesa A. "Differential expression in RNA-seq: A matter of depth", Genome Research, 2011, 21(12):2213-23.

[16] McCullagh P, and Nelder J. "Generalized Linear Models", CHAPMAN & HALL/CRC, 1989.

[17] Friedman J, Hastie T, and Tibshirani R."Regularization Paths for Generalized Linear Models via Coordinate Descent.", J Stat Softw, 2012, 33 (1):1-22.

[18] Lister R, Pelizzola, M, Kida Y, Hawkins R, Nery J, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, Downes M, Yu R, Stewart R, Ren B, Thomson J, Evans R, and Ecker J. "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells", Nature, 2011, 471: 68 - 73.

[19] Reinke AW, Baek J, Ashenberg O, Keating AE. "Networks of bZIP protein-protein interactions diversified over a billion years of evolution", 2013, 340(6133):730-4.

[20] Lunardi A, Di Minin G, Provero P, Dal Ferro M, Carotti M, Del Sal, G, Collavin L. "A Genome-Scale Protein Interaction Profile of Drosophila P53 Uncovers Additional Nodes of the Human P53 Network", Proc. Natl. Acad. Sci. 2010, 107(14):6322-7.

[21] Yang A, Kaghad M, Wang Y, Gillett E, Fleming MD, Dtsch V, Andrews NC, Caput D, McKeon F. "p63, a p53 homolog at 3q27-29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities", Mol. Cell. 1998, 2 (3): 30516.

[22] Dong X., Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein G, Guig R, Birney E and Weng Z. "Modeling gene expression using chromatin features in various cellular contexts", Genome Biology. 2012,13:R53.

[23] Wang C, Zhang MQ, and Zhang Z. "Computational identification of active enhancers in model organisms", Genomics Proteomics Bioinformatics, 2013, 11(3):142-50.