ChIP-seq data plays an important role in a cytosine-based DNA methylation prediction model

Jie Lv[1], Yan Zhang[1,*], Yunfeng Qi[1], Hongbo Liu[1], Jiang Zhu[1], Jianzhong Su[1] and Ruijie Zhang[1]
[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China

Abstract

DNA methylation was found previously related with histone modifications. The relationship among DNA methylation and histone modifications is potentially identifiable by integrating ChIP-seq and methylation data. However, little has been addressed on this issue in literature. Predicting DNA methylation can largely avoid the high cost of bisulfite-converted DNA experiments and drawbacks of microarray-based approaches. The most important, biological studies can be designed in a more economical manner. In this study, we found the DNA methylation was strongly influenced by surrounding combinatorial enriched ChIP-seq derived features in genome-wide scale. As an application, a cytosine-based methylation prediction model is proposed to predict the methylation status. As a result, we found Lymph-specific genes were distinct from other kinds of genes around Transcription Start Sites, which confirmed that our model is tissue-specific.

1. Introduction

DNA methylation is highly related with various biological processes in higher organism including development and genomic imprinting, and shows certain unclear methylation patterns [1]. Though DNA methylation has an apparent regulatory role, the measurement is handicapped by biological technologies [2]. Especially, precise DNA methylation globally is invisible to inaccurate microarray-based experiments and threatened by high cost of bisulfite-converted DNA experiments.

The histone modifications such as methylation and acetylation also take part in the transcriptional regulation. Different combinations of histone modifications may result in nearly dissimilar DNA methylation status [3], which enlightens the relevance of histone modifications in methylation studies. Sequencing histone modifications using ChIP-seq technology provides an opportunity in precisely prediction of DNA methylation. Actually, ChIP-seq experiment is superior to direct methylation probing experiments with respect to experimental cost, precision and inborn genome-wide scale.

Recently, Fan et al. used a window method with fixed size to further improve CpG island (CGI) methylation prediction using ChIP-seq data, especially four histone modifications [4]. However, prior studies [5, 6] including Fan et al. failed to consider the following aspects. On the one hand, the window method failed to capture the “true” binding sites of histone modifications in Fan et al.’s model. On the other hand, all previous studies were laboriously based or partly based on hundreds of epigenomic features such as DNA sequence patterns. These features are of little use in improving prediction performance, compared with ChIP-seq derived features. Moreover, prior studies only focused on CGIs defined by classical criteria. Here, our prediction targets are extended beyond CGIs. With respect to biological experiment, the microarray-based approach is biased towards GC contents [7]. Though straightforward, they are not designed in genome-wide scale indeed [7]. Compared with direct probing, ChIP-seq data, we suppose, is superior to direct microarray-based experiments with respect to both cost and accuracy in methylation prediction. The prediction precision is improved by our model, which is better than previous “high or low” methylation models.

In this study, the derived precise relationship between the centeredness of given cytosines relative to “true” binding sites and cytosine methylation were applied to infer precise region methylation status. Specific region enriched with tags is referred to the precise binding location. The Cytosine-based Region Methylation Prediction Model (CRMPM) integrating both epigenetic and epigenomic features was established to torture the methylation status in human CD4+ T cells. When restricting the predicting targets to CGIs, our model outperformed Fan et al. on the experimental human brain data. When applying CRMPM to tissue-specific genes, the methylation status of Lymph-specific genes were found to be

* Corresponding author.
E-mail addresses: yanyou1225@yahoo.com.cn (Y. Zhang), lvjielvjie@yahoo.com.cn (J. Lv)

978-0-7695-3735-1/09 $25.00 © 2009 IEEE
DOI 10.1109/FSKD.2009.708

IEEE computer society

distinct from other tissue-specific and housekeeping genes, which confirmed CRMPM was tissue-specific. Software and sample datasets are freely available at http://www.softpedia.com/get/Others/Home-education/PDMGE.shtml.

2. MATERIALS AND METHODS

2.1. Dataset

The DNA methylation data in CD4+ T cells was derived from the Human Epigenome Project (HEP) (http://www.sanger.ac.uk/PostGenomics/epigenome/). The histone modification data were taken from Barski et al, in which they used ChIP-seq experiment to sequence 38 kinds of histone methylation and acetylation modifications in human CD4+ T cell (http://dir.nhlbi.nih.gov/Papers/lmi/epigenomes/) [8].

The genomic features we used included the CGI annotation predicted by CpGcluster (http://bioinfo2.ugr.es/CpGcluster/), the repeat data downloaded from UCSC, and the promoters defined by [4]. In addition, ChIP-seq derived epigenomic features of CTCF, PolII were also obtained from Barski et al.

The dataset of Rollins et al. (http://epigenomics.cu-genome.org/html/meth_landscape/) was used to validate our model [9], which contains 4240 human brain unmethylated domains (considered as low methylation status) and 3518 methylated domains (considered as high methylation status). In addition, tissue-specific (http://bioinfo.wilmer.jhu.edu/tiger/) and housekeeping genes (http://www.cgen.com/supp_info/) were also used for validation. The methylation status of 344 Lymph-specific genes, employing 236 Brain, 309 Liver and 254 Placenta-specific genes as positive control genes, and 555 housekeeping genes as negative control genes were predicted.

2.2. Model construction

To study the relationship between DNA methylation and diverse ChIP-seq data, we used MACS, an enrichment tool to gain remapped ChIP-seq profiles with both high true positive and low false positive rate [10]. In Zhang et al.'s model [10], ChIP-seq tags enriched within a given region than would be expected are clustered to form a "true" binding site. As a basis of our model, the ChIP-seq enrichment degrees (CEDs) for cytosines were calculated by linear mapping and normalization. The workflow of illustrating the CED calculation procedure is shown in Figure 1.

Numerical regression and discrete classification were both used for CRMPM.

In numerical case, linear regression was used to precisely estimate the probabilistic cytosine methylation status, coefficients in eq. (1) were estimated from the cytosine data.

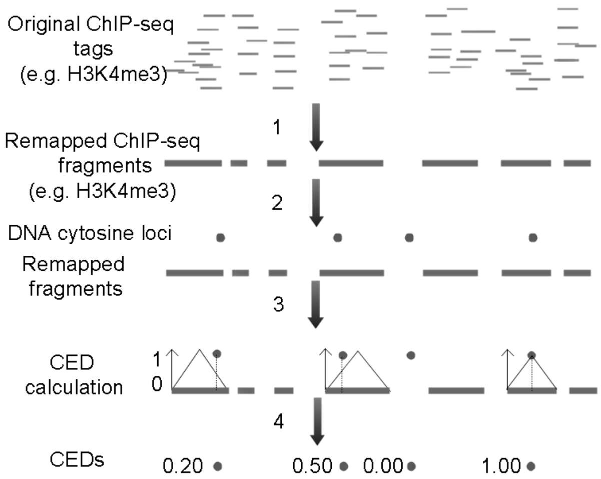
Original ChIP-seq tags (e.g. H3K4me3)
1
Remapped ChIP-seq fragments (e.g. H3K4me3)
2
DNA cytosine loci
Remapped fragments
3
CED calculation
1
0
4
CEDs
0.20
0.50
0.00
1.00

Fig. 1. The workflow of the ChIP-seq enrichment degree (CED) for cytosines calculation procedure. Enriched ChIP-seq fragments are marked by bars. (1) "True" binding sites for histone modifications or other proteins are remapped onto the genome. (2) Cytosine loci correlate with remapped tags. (3) CEDs are calculated by linear mapping and are normalized to interval of [0.00,1.00]. (4) The numerical CEDs for cytosines are calculated. We take four representing loci for instances. Three cytosines are associated with their overlapped enriched fragments, the relative position is 0.10, 0.50, 0.00 and 1.00, respectively. 3rd locus is considered free of enriched fragments, for the corresponding CED is 0.00. We suppose that 4th cytosine is influenced by trimethylation modification of lysine 4 in histone H3 most and 3rd cytosine is independent of trimethylation modification of lysine 4 in histone H3.

$$M_i = \beta_0 + \sum_{j=1}^{p} \beta_i CED_{ij} \qquad (1)$$

Let M_i be the numerical methylation status from HEP data, the range of M_i is 0.00 to 1.00 by averaging methylation status of redundant cytosines, and CED_{ij} be the CED for feature j in cytosine i. We filtered the features with p>0.01 (t-test). The cytosines with all zero CEDs were also filtered. After filtering, N is actually 16,685 in linear model and p is 21. The discrete epigenomic features were coded to adapt to eq. (1) according to their general influences upon DNA methylation (CGI: 0, Promoter: 0, Repeat: 1, non-CGI: 1, non-Promoter: 1, non-Repeat: 0).

The numerical region methylation status was defined in eq. (2).

$$RM_k = \frac{1}{n_k}\sum_{i=1}^{n_k} M_i \qquad (2)$$

Let RM_k denote the numerical methylation status of region k and n_k be the cytosine number in RM_k. RM_k was determined by averaging cytosine loci in region k.

To compare with previous models, bimodal status discretized from numerical status were required. In discrete case, the methylation status was divided into high and low status, which was supported by our model. Logistic regression classifier in Weka with default parameters was used, while simply by weighting the discrete class labels by extending Weka package. Let RM_k denote the illustrative methylation status, for region k. The definition of n_k was the same as eq. (2). RM_k was determined simply by weighing two class labels.

Low methylation status was treated as the positive class label in discrete case as other models. Standard Pearson correlation coefficient (PCC) was calculated in numerical case. In addition, sensitivity (Sn), specificity (Sp), accuracy (ACC), error rate (ER), precision (PR) and correlation coefficient (CC) were calculated in the usual way.

Analysis of variance (ANOVA) was carried out to evaluate the significance among multiple types of genes, and Tukey test was used to control pairwise multiple comparisons.

3. Results

3.1. ChIP-seq derived CEDs are highly related with cytosine methylation

As to test the predictability of cytosine methylation from ChIP-seq data, especially from various histone marks, the cytosine methylation was correlated with CEDs of epigenomic and epigenetic features. As a result, the PCC between the experimental cytosine methylation status and feature combinations was 0.7980 ($p<10^{-6}$) for both N=31,237 for original data and 16,685 for filtered data. The extremely significant features (p<0.01) are extracted as a basis for further prediction. These informative features were considered to be jointly associated with DNA methylation. In discrete case, the most discriminating low and high cutoffs had to be fixed. The cutoffs of 0.07 and 0.97 were supposed to be the ideal low and high cutoff, respectively, by the loci with the highest 1% average PR and with the maximum cytosines. Under the ideal cutoffs, the Average PR and ER of 10-fold cross-validation (CV) for bimodal classes achieved 0.9426 and 2.24%, respectively, indicative of discrete CEDs were informative in our locus-based model.

3.2. Validation by human brain methylation data

Generally, predicting model comprising ChIP-seq data are better than lacking. The only up-to-date pioneering model making use of ChIP-seq data was Fan et al.'s. Their model was superior to previous ones including Bock et al.'s, details saw [4]. Hence, model comparison was performed only with Fan et al. To this end, we investigated whether cytosine extraction with different cutoffs around domain center impacted on the validation performance. To faithfully do a comparison with Fan et al., the cutoff was determined as 0.375 (492 CGI domains), as it implied the closest number to Fan et al. (493 domains in their study). The cytosines in domains were extracted according to the following cutoff: normalized distance of 0.125 to domain center in either side. The benchmark results are listed in Table 1. We observed that our model (p=21) were superior to Fan et al. except Sn. With regards to Sn, 5 out of 20 cutoffs in this paper are superior to Fan et al. The CC and ACC were significantly better than pervious studies.

Table 1. The benchmark results for full model, and Fan et al.'s model, respectively.

Model	p	Sn	Sp	CC	ACC
CRMPM	21	0.822	0.996	0.817	0.900
Fan et al.	39	0.766	0.824	0.590	0.801

3.3. Methylation prediction of tissue-specific genes

To predict genome-wide gene methylation in CD4+ T cells, we selected cytosine loci within the region around TSSs with 5% of gene full length for both upstream and downstream to calculate the methylation status for each gene, as the subregion was ideally discriminative among tissue-specific and housekeeping genes. The length of subregions is suggested by our additionally larger range study (data not shown). We grouped the predicted methylation status for all kinds of genes. As an averaging result, the bar chart is shown in Fig. 2. Tukey test confirmed that the methylation status were ideally discriminating Lymph-specific genes from other types of genes (p<0.05). It is indicated that the tissue specificity of DNA methylation were identifiable by CRMPM. Notably in Fig. 2, Lymph-specific genes show higher methylation levels than housekeeping genes, yet slightly lower than other tissue-specific genes. The anti-correlation between promoter methylation and gene expression is also implied in the figure. The predicting results for Lymph-specific genes could be served as supporting information in subsequent laboratory experiment.

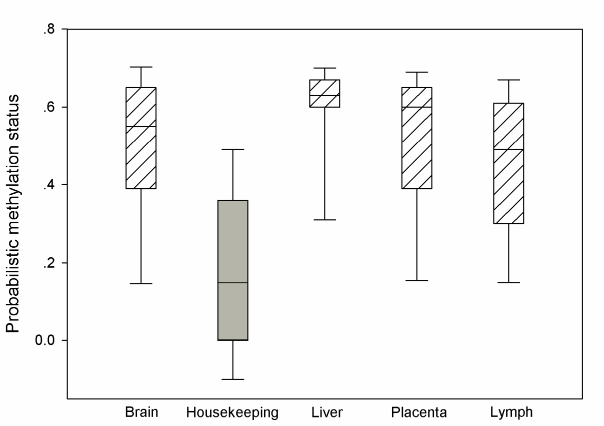

Fig. 2. The boxplot of predicted probabilistic gene methylation level in Lymph-specific genes with other tissue-specific genes as positive control and housekeeping genes as negative control. Outliers are not shown. Housekeeping genes marked by grey color are assumed to be low methylated, which are served as the negative control. Lymph genes are discriminating from other types of genes indicated by Tukey test.

3. Discussion

A locus-based genomic region methylation prediction model in human CD4+ T cells is proposed in this paper and is applied to predict the methylation status in tissue-specific and housekeeping genes. Biological validation confirmed the capability of CRMPM. Gene methylation prediction also confirmed the idea of the anti-correlation of methylation and transcription.

We suppose the application of CRMPM should not be confined to CGIs but be adapted to genomic regions with varying GC contents. Obviously, methylation is influenced by GC contents. Unfortunately, we do not take GC contents into account in this study, as CpG-rich and CpG-poor regions are hard to discriminate by available CpG islands finding algorithms.

CRMPM is also applicable when one supplies another feature sets. For studies of cancer epigenetics, the DNA methylation status in aberrant cells could also be inferred by corresponding histone modification profiles. However, the validation is not carried out owing to lack of appropriate experimental methylation and histone modifications data in cancer.

As the available ChIP-seq data is accumulating at a high speed, methylation prediction in other species should be applicable when appropriate, which would definitely improve the understanding of epigenetic regulation in other species.

4. Acknowledgements

This work was supported in part by the Natural Science Foundation of Heilongjiang Province (Grant Nos.D2007-35), the Heilongjiang Province Department of Education Outstanding Overseas Scientist grant (Grant Nos. 1152hq28), and the Innovation and Technology special Fund for researchers of Harbin (Grant Nos. RC2007LX003004).

5. References

[1] A. Bird, "DNA methylation patterns and epigenetic memory," Genes Dev, vol. 16, pp. 6-21, Jan 1 2002.

[2] B. E. Bernstein, A. Meissner, and E. S. Lander, "The mammalian epigenome," Cell, vol. 128, pp. 669-81, Feb 23 2007.

[3] R. Margueron, P. Trojer, and D. Reinberg, "The key to development: interpreting the histone code?," Curr Opin Genet Dev, vol. 15, pp. 163-76, Apr 2005.

[4] S. Fan, M. Q. Zhang, and X. Zhang, "Histone methylation marks play important roles in predicting the methylation status of CpG islands," Biochem Biophys Res Commun, vol. 374, pp. 559-64, Sep 26 2008.

[5] C. Bock, M. Paulsen, et al., "CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure," PLoS Genet, vol. 2, p. e26, Mar 2006.

[6] F. Fang, S. Fan, X. Zhang, and M. Q. Zhang, "Predicting methylation status of CpG islands in the human brain," Bioinformatics, vol. 22, pp. 2204-9, Sep 15 2006.

[7] V. K. Rakyan, T. A. Down, et al., "An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs)," Genome Res, vol. 18, pp. 1518-29, Sep 2008.

[8] A. Barski, S. Cuddapah, et al., "High-resolution profiling of histone methylations in the human genome," Cell, vol. 129, pp. 823-37, May 18 2007.

[9] R. A. Rollins, F. Haghighi, et al., "Large-scale structure of genomic methylation patterns," Genome Res, vol. 16, pp. 157-63, Feb 2006.

[10] Y. Zhang, T. Liu, et al., "Model-based Analysis of ChIP-Seq (MACS)," Genome Biol, vol. 9, p. R137, Sep 17 2008.

Proceedings

The Sixth International Conference on Fuzzy Systems and Knowledge Discovery

FSKD 2009

Volume 5

Proceedings

The Sixth International Conference on Fuzzy Systems and Knowledge Discovery

Volume 5

Tianjin, China
14–16 August 2009

Sponsored by
Tianjin University of Technology

Editors
Yixin Chen, Hepu Deng, Degan Zhang, and Yingyuan Xiao

Los Alamitos, California
Washington • Tokyo

Published by the IEEE Computer Society
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314

IEEE Computer Society Order Number P3735
BMS Part Number: CFP09FSK-PRT
Library of Congress Number 2009906464
ISBN 978-0-7695-3735-1

ISBN 978-0-7695-3735-1
90000
9 780769 537351

Vol. 5

Sixth International Conference on
Fuzzy Systems and Knowledge Discovery (FSKD 2009)

Sixth International Conference on

Fuzzy Systems and Knowledge Discovery

Vol. 5

FSKD 2009

Tianjin, China • 14-16 August 2009

*Edited by Yixin Chen, Hepu Deng,
Degan Zhang, and Yingyuan Xiao*