A quantitative analysis about the relative relationship of histone modifications and transcription factor binding to chromatin accessibility

Abstract

It is a consensus that histone modifications and the binding of transcription factors exert a significant impact on the "openness" of chromatin. In this study, We present a quantitative analysis of the genome-wide relationship between chromatin features and chromatin accessibility. We found that these features show distinct preference to localize in open chromatin. In order to elucidate the exact relationship, we derived quantitative models to directly predict the "openness" of chromatin using histone modification features and transcription factor binding features respectively. We show that both these two types of features are highly predictive for chromatin accessibility in a statistical viewpoint. Moreover, our results indicate that these features are highly redundant and only a small number of features can achieve a very high predictive power. Our study provides new insights into the combinatorial effects of different chromatin features to chromatin accessibility.

Key words:

Chromatin accessibility, histone modifications, transcription factor binding, regression analysis

Abbreviations:

HM: histone modification

TFBS: transcription factor binding site

SVR: support vector regression

SCC: spearman correlation coefficient

1 Introduction

In eukaryotes, DNA is organized into chains of nucleosomes, which consists of about 146bp of DNA wrapped around an octamer of four types of histones [1]. The packaging of chromatin into nucleosomes provides a repressive environment for many DNA-binding proteins and plays a important role on the regulation of transcription [2]. However, some domains in chromatin are depleted of nucleosomes and exhibit highly accessible structure. These nucleosome-free regions are super sensitive to the cleavage of Dnase 1 [3]and are known as Dnase 1 hypersensitive sites (DHSs). They are found predominantly in many active genes and cis-regulatory elements [4]. The dynamic alterations of "openness" in chromatin play a import role in many biological processes, including transcription [5], replication [2]and differentiation [10].

Traditionally, the experimental techniques of choice to discover the Dnase 1 hypersensitive sites are Southern blots [6]. However, this low-throughput method is not able to study large chromosomal regions at a time and can't represent the "openness" of chromatin in a quantitative manner. The meaning of differential accessibility is unknown, but may reflect some important biological phenomenon such as histone modifications and protein occupation [14]. Until now genome-wide quantitative analysis of the relationship between Dnase 1 hypersensitive sites and chromatin features is rare. By taking advantage of the abundant datasets of the ENCODE project [12] ,we analyzed genome-wide localization of Dnase 1 hypersensitive sites and 33 chromatin features in H1hese cell line. All datasets are generated by recently developed genome-wide experimental techniques ,such as Chip-seq [7, 8]and Dnase-seq [9].

It is generally accepted that histone modifications and the binding of transcription factors are two main effectors for the "openness" of chromatin. Previous studies have shown that histone modifications and transcription factors tend to occur near or just in the DHS [11, 14]. Recently, two studies, one in K562 cell line and the other in Drosophila embryonic cells, have demonstrated that transcription factor binding sites and the chromatin accessibility are highly correlated with each other [10, 11]. Although these studies have provided important information, So far, quantitative analysis about the combinatorial effects of different chromatin features is still absent. As an extension, we build support vector regression (SVR) models to directly predict the "openness" of chromatin using chromatin features. Our work not only confirms these previous findings, but also indicates that both histone modification features and transcription fac-

tor features are predictive for chromatin accessibility with high accuracy and these chromatin features are highly redundant.

2 Materials and Methods

2.1 Datasets.

All datasets are from ENCODE project, which aims to build a comprehensive list of functional elements in the human genome [12]. The 10 histone modifications and 23 transcription factor binding sites were quantified using Chip-seq [7, 8]. The chromatin accessibility dataset was measured using Dnase-seq [9]. Each dataset includes the genome-wide signals and regions of signal statistically enrichment (peaks). Peaks can be viewed as locations of chromatin features and Dnase 1 hypersensitive sites respectively.

2.2 Mapping HM and TFBS peaks on the Dnase 1 hypersensitive sites

We obtained genomic locations of 33 chromatin features (Chip-seq peaks), including 582489 HM peaks and 443217 TF peaks. For each feature, we mapped the peaks on the genome. The presence or absence of chromatin features within accessible chromatin was decided by overlap or non-overlap. If there was any amount of overlap within accessible chromatin (Dnase-seq peaks), we counted as a presence [13]. Then, we calculated the percentage of how many peaks occurring in the Dnase 1 hypersensitive sites for each feature.

2.3 Supervised learning methods for chromatin accessibility prediction

To investigate the quantitative relationship between chromatin accessibility and these chromatin features, we constructed support vector regression (SVR) models for histone modifications and transcription factor binding features respectively. Concretely, in every DHS, we calculated the max Dnase-Seq signal and the max Chip-Seq signal of every chromatin feature. Then, SVR model was built to predict the chromatin accessibility using signals of these chromatin features. SVR is a machine learning algorithm based on statistical theory for regression problems [16, 19]. We implemented this algorithm using the "e1071"R package [15].

In order to reduce the computation cost, we randomly select 5000 DHS for our samples. The sample size is enough to represent the entire datasets (AddFigure). We use the 10 fold cross-validation method to evaluate the prediction power. Specifically, we randomly split our whole dataset into 10 equal size subsets. a single subset is treated as the validation data for testing the model, and the remaining 9 subsets are used as training data. This precess is repeated 10 times and each subset can only be used once as the validation data. After that, we combined the results and calculated the spearman correlation coefficients (SCC) between predicted signal and actual Dnase-seq signal. The SCC can be viewed as the prediction power. And the square of SCC (coefficient of determination) can can be viewed as the proportion of the variation in chromatin accessibility that can be explained by the model.

2.4 Analysis of the importance and combinatorial effects of chromatin features

To estimate which feature exhibits the maximal prediction power, we predicted the chromatin accessibility using only one feature. And to investigate whether HM features and TFBS features are redundant, we next predicted the 'openness'of chromatin using all features. We also explored the combinatorial effects of these features. All possible one-feature (C_{33}^1), two-features (C_{33}^2) and three-features (C_{33}^3) models were evaluated by their performance.

2.5 Model comparison analysis

Instead of SVR algorithm, we also explore the quantitative relationship between chromatin features and chromatin accessibility with liner regression model. Similarly, HM features, TF features and HM+TF feature combinations are applied into linear regression model respectively. The spearman correlation coefficients of the predicted signals and the actual Dnase-seq signal are calculated and compared with the SVR models. In order to figure out whether the max signals or the average signals of chromatin features in the Dnase-seq peaks exhibit largest prediction power, we also applied these models with average signals of chromatin features in this region.

3 Results

3.1 The localization preference of chromatin features

We analysed genome-wide localization of 33 Chip-seq profiles in the human embryonic stem cell line (H1hesc) from ENCODE project [12], including 10 histone modifications, and 23 transcription factor binding sites. For each profile, we mapped the peaks of Chip-seq dataset in the Dnase 1 hypersensitive sites (see Materials and methods). **Figure1** shows the percentage of how many peaks within the accessible chromatin for each feature. We observed that different chromatin feature exhibits different preference to chromatin accessibility. For histone modifications, H3k4me3 exerts the largest preference of accessible chromatin. 82.2% H3k4me3 peaks located in DHS. On the contrary, most H3k9me3 occured out of DHS (93.7%), which indicates that H3k9me3 is associate with heterochromatin [28]. Compared to histone modifications, a majority of transcription factors tend to bind on accessible chromatin, which suggests that the process of transcription requires a open chromatin structure [20]. The mean percentage of TFs locating in DHS is 60.5%, higher than histone modifications (45.1%).

3.2 Predicting chromatin accessibility using histone modifications.

In order to examine the quantitative relationship between chromatin accessibility and histone modification features in a combinatorial manner, we constructed SVR model to predict the "openness"of chromatin using all histone features. From **Figure2 (a)** we can see that the predicted signals and the actual Dnase-Seq signals are highly correlated with each other. The spearman correlation coefficient (SCC=0.70) can be viewed as the prediction power. This analysis suggests that histone modification features explain about 50% variance of chromatin accessibility.

We next examined the prediction power for every histone feature. Figure2 (b) shows that H3k4me3, H3k4me2 and H3k9ac exhibit the most important effects to chromatin accessibility (SCC = 0.58, 0.58, 0.57 respectively). These histone modifications are generally enriched in the promoters of expressed genes [29] and the open chromatin structure plays a important role in regulating the complex transcription process. On the other hand, H3k27me3, H3k9me3 and H3k36me3 exhibit least prediction powers (SCC= 0.33, 0.30, 0.21 respectively),

which suggests that these modifications are associated with heterochromatin [23, 26]. Interestingly, H3k27ac and H4k20me1, which are the most predictive histone modifications for gene expression levels [33], are not the most important features associated with chromatin accessibility.

3.3 Predicting chromatin accessibility using TFBS features

Previous studies have shown that transcription factors tend to bind on open chromatin and they are highly correlated with each other [10, 11]. To investigate the the quantitative relationship of the binding of transcription factors and chromatin accessibility in a combinatorial manner, we next applied our SVR model to all TFBS features. As shown in Figure3 (a), the TF model achieves a correlation of 0.73 which is a little higher than HM model. These TF features can explain more than 50% variance of chromatin accessibility.

For the prediction power of particular TF feature, there is a difference with histone modifications that most transcription factors exhibit important effects to chromatin accessibility (Figure3(b)). This is consistent with their functions because transcription factors are directly control the complex transcription process [30] which requires a open chromatin environment . However, a small group of feature exhibit lower prediction powers, such as ZNF274, SUZ12 and CTCF (SCC=0.31, 0.38, 0.42 respectively). ZNF274 and SUZ12 are known to be transcriptional repressors [27, 31]. CTCF has many roles, such as transcriptional repression, insulator function, and imprinting genetic information [32]. These factors are not so important to contribute to the "openness" of chromatin.

3.4 Chromatin features are highly redundant to chromatin accessibility

The above analyses suggest that both histone modification features and transcription factor features are predictive for chromatin accessibility with high accuracy. So there is a question that whether the prediction power will increase if we use all these features. To address this question, we directly predict the "openness" of chromatin using all features. As shown in Figure4 (a), the prediction power (SCC=0.77) is only a little higher than only using TF features which indicates that these two types of features are highly redundant. To check the importance of different features and their combinatorial effects, we build models

with all possible combinations of one to three feature (Figure4 (b)). Focusing on the three-features combinations (5456 models), we found that the least prediction power combinations (SCC=0.45) can also achieve more than 58% prediction power of the full model. And there are 137 combinations achieve more than 90% prediction power of the full one. These analyses indicate that most of these features are highly redundant for chromatin accessibility.

By examining the 137 high prediction power combinations, we found that five chromatin features , SUZ12, SIN3A, H3k4me3, H3k9ac, GTF2F1, are significantly enriched ($p < 0.01$) in this set of models. Interestingly, all these features have high prediction powers in the one-feature models except SUZ12. This may due to the lower redundancy of SUZ12, which is supported by the finding that the correlation of SUZ12 levels with the other four features are 0.14, 0.21, 0.26 and 0.12. SUZ12 is a part of Polycomb Repressive Complex 2 (PRC2) and may be involved in chromatin silencing with non-coding RNA[27]. The mechanisms of how SUZ12 impact chromatin structure is unknown, however, it may be different with other features.

3.5 Comparison with other models

In this study, we choose the SVR algorithm and the max signal in every region modeling the relationship between chromatin features and chromatin accessibility. Generally, The SVR algorithm is a nonlinear regression method. We also have explored their relationship using liner regression model and the average signal in every region. As shown in Table1, prediction power of models using average signal are significantly lower than the max signal models. And in either situation, the SVR models exhibit higher prediction power than liner models. Our results indicate that the "openness" of chromatin are determined by the max signal of features and their relationships are assumed as a non-liner relevance.

4 Discussion

In this work, we present a quantitative analysis about the relationship of histone modifications and the binding of transcription factors to chromatin accessibility. We first examined the percentage of feature peaks within DHS in H1hesc cell line. We find that different chromatin features show different location preference in DHS. This may duo to the particular function of different

chromatin features. Robert E. Thurman et al have done similar analysis in K562 cell line [12]for TF features. There is a big difference that the percentage of transcription factors within DHS seems significant higher than in the H1hesc cell line. This may because in order to maintain the 'stemness' state, most genes are repressed in the stem cell compared to the cancer cell line K562. This phenomenon means that the degree to what extent chromatin features occur in accessible chromatin may differ according to different cellular circumstances.

Our results indicate that both HM features and TF features account for nearly or more than 50% variation of chromatin accessibility in H1hesc cell line. For histone marks, many activators of gene expression exhibit important impact on the "openness" chromatin, such as H3k4me [23] and histone acetylations [24]. The hallmarks of repressed genes, such as H3k9me3 [23], have lower prediction powers. Unexpectedly, the transcription elongation mark H3k36me3 [25] shows the least prediction power. This is consistent with the viewpoint of an recently published paper. Sophie Chantalat et al [26] argues that H3k36me3 is associated with constitutive and facultative heterochromatin. For TF features, the majority of TFs shows an important impact on chromatin accessibility except some transcriptional repressor, such as ZNF274 and SUZ12. This indicates that the complex transcription process requires open chromatin environment [20].

It is generally accepted that cellular factors regulate the complex dynamic change of chromatin structure in a collective way. We have shown that these features are highly redundant to predict chromatin accessibility and only a small subgroup of features can achieve a very high prediction power. However, the mechanism of how these features cooperatively impact the openness of chromatin is still unclear and we must note that our analysis could not reveal the 'cause' or 'consequence' relationship of HM and TF features to chromatin accessibility. Histone modifications play a important role in creating and maintaining the accessible chromatin environment [22] and may act as docking sites for transcription factors [17]. Some pioneer factors tend to bind on the genome and create a accessible site, such as FoxA1 [21] which is the best known pioneer factor. Then, more transcription factors tend to bind on the opening site and the Dnase1 hypersensitive site is created. As an extension, future work could explore the mechanisms of how these features cooperatively regulate open chromatin structure and their causal relationships with additional data.

Figure Legends

Figure1 . The percentage of histone modification and transcription factor binding features within accessible chromatin. The black circle and blue triangle identify histone modification features and TFBS features respectively. The two red lines represent the mean percentages for HM and TF respectively.

Figure2. The prediction power of chromatin accessibility with histione features. (a) Scatterplot of predicted versus experimentally measured Dnase-seq signals using all histone features. The black line represents the liner fit between predicted and measured signal (SCC, spearman correlation coefficient). (b) Prediction power of the SVR models using only one particular histone feature.

Figure3. The prediction power of chromatin accessibility with TF features. (a) Scatterplot of predicted versus experimentally measured Dnase-seq signals using all TF features. The black line represents the liner fit between predicted and measured signal (SCC, spearman correlation coefficient). (b) Prediction power of the SVR models using only one particular TF feature.

Figure4. Redundancy of histone features and TF features. (a) Scatterplot of predicted versus experimentally measured Dnase-seq signals using all histone and TF features. (SCC, spearman correlation coefficient). (b) Comparison of prediction power between all possible one-feature, two-features, three-features models and the full model in H1hesc.

Table1. Comparison of prediction powers with different models. The prediction power is represented as Spearman correlation coefficient of predicted signal and the actual Dnase-seq signal. LM: linear regression model. Max_sig: max signal. Avg_sig: average signal.

Supplementary materials

AddFigure. The prediction powers with different sample sizes. In each sample size, we sampled 500 times. We can see that the prediction power increases with the rising of sample size. However, when the sample size reaches 2000, the prediction power increases gently. The mean prediction power of 5000 sample size is 0.774 ± 0.006 .

References

[1] Luger, Karolin, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. "Crystal structure of the nucleosome core particle at 2.8 Å resolution." Nature 389, no. 6648 (1997): 251-260.

[2] Anderson, J. D., and J. Widom. "Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites." Journal of molecular biology 296, no. 4 (2000): 979-987.

[3] Dingwall, Colin, George P. Lomonossoff, and Ronald A. Laskey. "High sequence specificity of micrococcal nuclease." Nucleic acids research 9, no. 12 (1981): 2659-2674.

[4] Gross, David S., and William T. Garrard. "Nuclease hypersensitive sites in chromatin." Annual review of biochemistry 57, no. 1 (1988): 159-197.

[5] Cockerill, Peter N. "Structure and function of active chromatin and DNase I hypersensitive sites." FEBS Journal 278, no. 13 (2011): 2182-2210.

[6] Lu, Qianjin, and Bruce Richardson. "DNaseI hypersensitivity analysis of chromatin structure." In Epigenetics Protocols, pp. 77-86. Humana Press, 2004.

[7] Park, Peter J. "ChIP–seq: advantages and challenges of a maturing technology." Nature Reviews Genetics 10, no. 10 (2009): 669-680.

[8] Mardis, Elaine R. "ChIP-seq: welcome to the new frontier." Nature methods 4, no. 8 (2007): 613-613.

[9] Song, Lingyun, and Gregory E. Crawford. "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells." Cold Spring Harbor Protocols 2010, no. 2 (2010): pdb-prot5384.

[10] Li, Xiao-Yong, Sean Thomas, Peter J. Sabo, Michael B. Eisen, John A. Stamatoyannopoulos, and Mark D. Biggin. "The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding." Genome Biol 12, no. 4 (2011): R34.

[11] Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield et al. "The accessible chromatin landscape of the human genome." Nature 489, no. 7414 (2012): 75-82.

[12] Dunham, Ian, Ewan Birney, Bryan R. Lajoie, Amartya Sanyal, Xianjun Dong, Melissa Greven, Xinying Lin et al. "An integrated encyclopedia of DNA elements in the human genome." (2012).

[13] Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield et al. "The accessible chromatin landscape of the human genome." Nature 489, no. 7414 (2012): 75-82.

[14] Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. "High-resolution mapping and characterization of open chromatin across the genome." Cell 132, no. 2 (2008): 311-322.

[15] Dimitriadou, Evgenia, Kurt Hornik, Friedrich Leisch, David Meyer, and Andreas Weingessel. "Misc functions of the Department of Statistics (e1071), TU Wien." R package (2008): 1-5.

[16] Cristianini, Nello, and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[17] Bell, Oliver, Vijay K. Tiwari, Nicolas H. Thomä, and Dirk Schübeler. "Determinants and dynamics of genome accessibility." Nature Reviews Genetics 12, no. 8 (2011): 554-564.

[18] Taverna, Sean D., Haitao Li, Alexander J. Ruthenburg, C. David Allis, and Dinshaw J. Patel. "How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers." Nature structural & molecular biology 14, no. 11 (2007): 1025-1040.

[19] Cheng, Chao, Koon-Kiu Yan, Kevin Y. Yip, Joel Rozowsky, Roger Alexander, Chong Shou, and Mark Gerstein. "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets." Genome Biol 12, no. 2 (2011): R15.

[20] Sproul, Duncan, Nick Gilbert, and Wendy A. Bickmore. "The role of chromatin structure in regulating the expression of clustered genes." Nature Reviews Genetics 6, no. 10 (2005): 775-781.

[21] Cirillo, Lisa Ann, Frank Robert Lin, Isabel Cuesta, Dara Friedman, Michal Jarnik, and Kenneth S. Zaret. "Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4." Molecular cell 9, no. 2 (2002): 279-289.

[22] Marx, Jean. "Protein Tail Modification Opens Way for Gene Activity." Science 311, no. 5762 (2006): 757-757.

[23] Margueron, Raphaël, and Danny Reinberg. "Chromatin structure and the inheritance of epigenetic information." Nature Reviews Genetics 11, no. 4 (2010): 285-296.

[24] Turner, Bryan M. "Histone acetylation and an epigenetic code." Bioessays 22, no. 9 (2000): 836-845.

[25] Fuchs, Stephen M., R. Nicholas Laribee, and Brian D. Strahl. "Protein modifications in transcription elongation." Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1789, no. 1 (2009): 26-36.

[26] Chantalat, Sophie, Arnaud Depaux, Patrick Héry, Sophie Barral, Jean-Yves Thuret, Stefan Dimitrov, and Matthieu Gérard. "Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin." Genome Research 21, no. 9 (2011): 1426-1437.

[27] Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., ... & Chang, H. Y. (2007). Functional Demarcation of Active and Silent Chromatin Domains in Human< i> HOX</i> Loci by Noncoding RNAs. Cell, 129(7), 1311-1323.

[28] Bartkova, J., Moudry, P., Hodny, Z., Lukas, J., Meyts, R. D., & Bartek, J. (2011). Heterochromatin marks HP1γ, HP1α and H3K9me3, and DNA damage response activation in human testis development and germ cell tumours. International journal of andrology, 34(4pt2), e103-e113.

[29] Regha, Kakkad, Mathew A. Sloane, Ru Huang, Florian M. Pauler, Katarzyna E. Warczok, Balázs Melikant, Martin Radolf et al. "Active and repressive chromatin are interspersed without spreading in an imprinted

gene cluster in the mammalian genome." Molecular cell 27, no. 3 (2007): 353-366.

[30] Gill, G. (2001). Regulation of the initiation of eukaryotic transcription. Essays Biochem, 37, 33-43.

[31] Yano, K., Ueki, N., Oda, T., Seki, N., Masuho, Y., & Muramatsu, M. A. (2000). Identification and characterization of human ZNF274 cDNA, which encodes a novel kruppel-type zinc-finger protein having nucleolar targeting ability. Genomics, 65(1), 75.

[32] Dunn, K. L., & Davie, J. R. (2003). The many roles of the transcriptional regulator CTCF. Biochemistry and cell biology, 81(3), 161-167.

[33] Karlić, R., Chung, H. R., Lasserre, J., Vlahoviček, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. Proceedings of the National Academy of Sciences, 107(7), 2926-2931.