Prediction of Chromatin State Variability from Genomic Sequence

Luca Pinello[1,2], Jian Xu[2,3,4], Stuart H. Orkin[1,2,3,4], **Guo-Cheng Yuan**[1,2]

[1]Dana-Farber Cancer Institute, [2]Harvard University, [3]Boston Children's Hospital, [4]Howard Hughes Medical Institute

Background

In eukaryotic cells the genome is organized into chromatin. The accessibility of the chromatin varies from one cell-type to another. The resulting constraint on protein-DNA binding provides an important layer of gene regulation. Recent epigenomic studies have uncovered diverse classes of regulatory elements, many of which are located in the regions previously viewed as "junk" DNA, providing strong evidence that chromatin states play a critical role in mediating cell-type specific transcriptional activities. However, the mechanisms underlying the variation of chromatin states remain poorly understood.

We have investigated the role of DNA sequence in mediating the cross cell-type variability of chromatin states with the focus on the histone mark H3K27me3, which mediates cell-type specific gene silencing [1] and plays an important role in the maintenance of cell identity and lineage differentiation [2] [3] . While it is well known that H3K27me3 occupancy is highly enriched at GC rich DNA elements [4], here we focus on distal regions where its recruiting mechanism is less understood [5] .

Results

Genome-wide Characterization of H3K27me3 Plasticity

We obtained a ChIPseq dataset containing H3K27me3 in 19 human cell lines from the ENCODE consortium [6]. The raw-sequence reads data were normalized and mapped to non-overlapping bins of 200bp. The fluctuation of sequence reads can be approximately modeled by a Poisson distribution, which has the distinct property that the mean level is equal to the variance. This motivated us to use the index of dispersion (IOD) statistic to quantify H3K27me3 variability. The Poisson distribution correspond to an IOD value of 1. We selected the top 1% of bins with highest IOD values and referred to those as the most variable regions (MVR) (Figure 1, green dots), whereas the bottom 1% of bins were referred to as the least variable regions (LVR) (Figure 1, red dots).

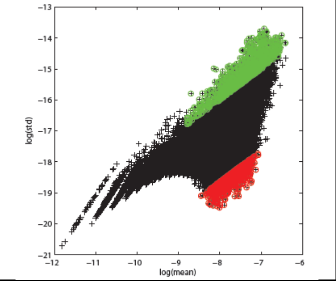

Figure 1. Definition of the MVR.

To test whether the variation of H3K27me3 was indeed associated with cell-type specific gene regulation, we investigated the correlation between the dynamic change of H3K27me3 occupancy and the expression levels of the neighboring genes. We found that, for most regions, there is a significant correlation between H3K27me3 density and gene expression level.

Prediction of H3K27me3 Plasticity from Genomic Sequences

The genome-wide distribution of H3K27me3 is regulated by both sequence dependent and independent mechanisms. On one hand, previous studies have identified a number of DNA sequence features associated with H3K27me3, including CpG islands [4], transcription factor sequence motifs [7] [8], short RNA hairpins [9], and lincRNA [10] [11] . On the other hand, existing H3K27me3 can be recognized by chromatin regulators thereby propagating in a self-enhancing manner. Previous studies have been focused on a specific cell-type, whereas to what extent the DNA sequence regulates the overall variability remains poorly understood. Of note, while the prediction of cell-type specific changes requires additional factors than sequence information, which is cell-type independent, it remains possible to predict the overall variability with accuracy as shown below.

We applied N-score, a computational method previously developed for nucleosome positioning prediction [12], to predict the location of MVRs from the underlying genomic sequences, using LVRs as negative control. We evaluated the model performance by cross-validation and obtained an accuracy of AUC = 0.82 (Figure 2). We then applied our model over running windows across the entire genome, and compared the predicted variability with ChIPseq data. The genome-wide correlation with experimental data is $\rho = 0.28$.

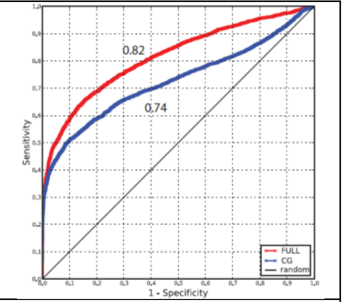

Figure 2. ROC for prediction of MVR locations from DNA sequence.

Distal MVRs Are Regulated by Cell-type Specific Transcription Factors

Next we focused on the MVRs in distal regions, which have recently been found to contain many important enhancer elements. We compared with a publicly available dataset of genome-wide enhancers in 9 ENCODE cell lines [13], and found that the distal MVRs are highly enriched with enhancers present in at least one cell line (p-value < 2.2E-16).

Compared to proximal MVRs, the distal MVRs tend to have lower mean and variance. Interestingly, the H3K27me3 density at distal MVRs appear to be bimodal: while the value is comparable to background level in most cell lines, it is significantly higher in one or two specific cell-types, suggesting an important role of cell-type specific regulators in their recruitment.

Since the distal MVRs are markedly cell-type specific, we searched for candidate TFs that may a role in Polycomb group (PcG) recruitment in cell-type specific manner. For each cell-type, we ranked the MVRs according to the z-score and selected the top ranking ones as the cell-type specific subset. We searched for known transcription factor (TF) motifs that are over-represented in each cell-type specific MVRs while using the rest as the background. For most cell-types, we were able to identify a small number motifs that are highly significantly over-represented.

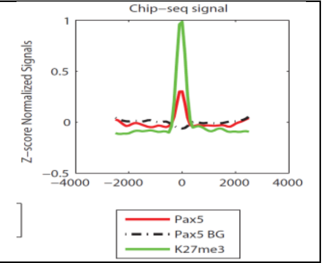

Figure 3. PAX5 is enriched in lymphoblastsoid-specific MVRs.

As an example, we found that the PAX5 motif is highly enriched in the lymphoblastoid cell lines (GM12878 and GM06990). Furthermore, the expression level of PAX5 is also higher in this cell-type than others, consist with the known role of PAX5 in B-cell development. Indeed, a role of Polycomb recruitment of PAX5 has previously been identified. To test whether PAX5 may facilitate PcG binding in this cell line, we tested its colocalization with H3K27me3 by using public ChIPseq data. We found that PAX5 and H3K27me3 indeed colocalize at these MVRs (Figure 3). Consistent with a gene silencing role, the target gene expressions are lower in these cell lines that the rest.

A Role of the TAL1 in Regulating H3K27me3 Recruitment in Erythroid Precursors

Next, we investigated whether the computational strategy discussed above may be useful for prediction of novel PcG recruiting factors in less well-characterized systems. In a recent study, we have characterized the genome-wide chromatin states in erythroid precursors (ProE) using primary human cell lines, and found that enhancer mediated gene activities are responsible for developmental-stage selection [14]. Using the same strategy as described above, we integrated our H3K27me3 ChIPseq data for ProE together with those obtained from ENCODE, and identified a subset of distal MVRs that are specific to ProE.

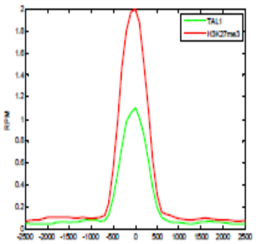

Figure 4. TAL1 is enriched in ProE specific MVRs.

In order to identify ProE-specific PcG recruiting factors, we searched for enriched TF motifs in the ProE-specific distal MVRs. One of the most enriched motifs corresponds to TAL1 (p-value = 5.6E-37). This is surprising because TAL1 is a well-characterized activator that is required for erythroid development. Although a possible role in repression has recently been suggested [15], a mechanistic understanding is still lacking. Our analysis suggests that TAL1 may play a role in PcG recruitment thereby repressing the target genes. We then examined the TAL1 ChIPseq data around the distal MVRs, and indeed found significant TAL1 binding signal (Figure 4). Furthermore, gene expression data analysis showed that the expression level of the target genes are expressed at a lower level compared to the overall TAL1 target genes. These results support a role of TAL1 in orchestrating PcG recruitment during erythroid development.

Conclusions

We have developed a systematic approach to investigate the mechanisms regulating chromatin state variability and applied it to H3K27me3. We found that the MVRs can be well-predicted by the underlying DNA sequences. Furthermore, the distal MVRs cannot be explained by GC content but are enriched for cell-type specific TF motifs. Using this approach, we found that the erythroid master regulator TAL1, which is commonly known as an activator, can also play a role in gene repression by targeted recruitment of Polycomb complexes. Our approach is generally applicable to other epigenetic marks.

1. Francis NJ, Kingston RE: **Mechanisms of transcriptional memory.** *Nat Rev Mol Cell Biol* 2001, **2:**409-421.
2. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, et al: **Polycomb complexes repress developmental regulators in murine embryonic stem cells.** *Nature* 2006, **441:**349-353.
3. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells.** *Cell* 2006, **125:**315-326.
4. Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B, Chi AS, Ku M, Bernstein BE: **GC-rich sequence elements recruit PRC2 in mammalian ES cells.** *PLoS Genet* 2010, **6:**e1001244.
5. Arnold P, Scholer A, Pachkov M, Balwierz PJ, Jorgensen H, Stadler MB, van Nimwegen E, Schubeler D: **Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting.** *Genome Res* 2013, **23:**60-73.
6. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
7. Liu Y, Shao Z, Yuan GC: **Prediction of Polycomb target genes in mouse embryonic stem cells.** *Genomics* 2010, **96:**17-26.
8. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, et al: **Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains.** *PLoS Genet* 2008, **4:**e1000242.
9. Kanhere A, Viiri K, Araujo CC, Rasaiyaah J, Bouwman RD, Whyte WA, Pereira CF, Brookes E, Walker K, Bell GW, et al: **Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2.** *Mol Cell* 2010, **38:**675-688.
10. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, et al: **Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.** *Proc Natl Acad Sci U S A* 2009, **106:**11667-11672.
11. Margueron R, Reinberg D: **Chromatin structure and the inheritance of epigenetic information.** *Nat Rev Genet* 2010, **11:**285-296.
12. Yuan GC, Liu JS: **Genomic sequence is highly predictive of local nucleosome depletion.** *PLoS Comput Biol* 2008, **4:**e13.
13. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473:**43-49.
14. Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA, Mikkola HK, Yuan GC, Orkin SH: **Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis.** *Dev Cell* 2012, **23:**796-811.
15. Van Handel B, Montel-Hagen A, Sasidharan R, Nakano H, Ferrari R, Boogerd CJ, Schredelseker J, Wang Y, Hunter S, Org T, et al: **Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium.** *Cell* 2012, **150:**590-605.