|  |
| --- |
| **Genome-wide prediction of chromatin accessibility based on gene expression** |

**Junyue Cao Jiayuan Guo Yu Liu**

**1 Scientific background**

The eukaryotic chromatin is tightly packaged into an array of nucleosomes, including a histone protein complex wrapped by DNA and separated by linker DNA[1]. The genome-wide positioning of nucleosomes affects the availability of DNA binding sites to transcription factors and thus regulates gene differential expression in development and disease [2]. The binding of transcription factors to promoters, enhancers and other regulatory elements in the genome also leads to nucleosome destabilization and results in open or accessible regions [3,4]. In this way, the chromatin accessibility reflects location and dynamics of gene regulatory elements.

Chromatin accessibility assays are used to isolate and identify either the accessible (e.g. DNase-seq [2][5,6], ATAC-seq [7]) or nucleosome protected DNA locations (Mnase-seq [8]). DNase-seq is one of the most widely used method for genome-wide chromatin accessibility profiling. In a typical experiment, the accessible DNA regions are cleaved by DNase I, followed by identification and quantification with high throughput sequencing [2]-[5,6]. The identified nucleosome-free genomic regions are characterized as DNase I hypersensitive sites (DHSs) [9]. It has been shown the DHSs closely correlates with the destabilization or loss of nucleosomes from regulatory elements during gene activation [10]. Despite several years of technique development, however, current DNase-seq method is still time-consuming and labor-intensive, and requires a lot of input materials. Furthermore, all chromatin accessibility assays normally need over 20 million reads for standard accessibility studies of the human genome[9], and leads to a high burden for sequencing cost. As a result, chromatin accessibility profiling via experimental procedure is limited in analyzing large number of samples [9].

Alternatively, chromatin accessibility can be inferred from gene expression, as both are tightly regulated by transcription factors binding to the cis-regulatory elements. The gene expression is measured by RNA sequencing, and is routinely collected from a broad spectrum of biological contexts, with more than >200,000 human samples in GEO database [11], in contrast to less than 7,000 DNase-seq data [12]. A high-performance prediction models for chromatin accessibility based on gene expression would help reveal the gene regulatory mechanism in greatly expanded biological contexts.

Here we tested the feasibility of predicting genome-wide chromatin accessibility based on gene expression. Unlike the previous studies that predict the chromatin accessibility in fixed 200 base pair bins across the genome [12], we combined Dnase-seq peaks from ~30 different human tissues from the Roadmap epigenomics project [13], and constructed the reference chromatin accessibility map for human cells. We then applied supervised machine learning to predicting the genome-wide chromatin accessibilities based on the gene expression within 1 Mb distance instead of the whole transcriptome. Here we show our prediction model can successfully reconstruct chromatin accessibility map in the test samples and provided a valuable tool for decoding gene regulatory elements in large scale analysis.

**2 Specific Aims:**

We will apply supervised machine learning (regularized linear regression, generalized additive model with smoothing spline , PCR and PLS regression) to predicting genome-wide chromatin accessibilities from transcriptome data of human tissues.

**3 Data Description:**

The DNase-seq peak data (aligned based on hg19) from around 50 tissues were downloaded from the Roadmap Epigenomics project [13]. The identified peaks for all tissues were combined, with overlapped peaks merged. The signal intensity for each DHS peak for each tissue were calculated by summing up the fold change value averaged between replicates within each peak, divided by the tissue’s total signal values and scaled by multiplying 1,000,000. The normalized signal was then log2 transformed after adding a pseudo-count 1. This transformed value represents the chromatin accessibility in each DHS peak.

The RNA-seq gene expression data (aligned based on hg19) from the same set of tissues were downloaded from the Roadmap Epigenomics project [13]. The FPKM (fragment per kilobase of exon per million mapped fragments) were then log2 transformed after adding a pseudo-count 1 and quantile normalized across samples. The RNA-seq samples in the test set were quantile normalized against the normalized samples in the training set.

**4 Proposed Approach and methods**

One challenge in this project is that we are trying to predict a high dimensional data (genome-wide chromatin accessibility, > 10,000 DHS sites) from another high dimensional data (transcriptome data, >30,000 genes). The conventional evaluation method (R square and MSE) for a single variable does not fit. Instead, we will compare different models based on cross-cell-type correlation: the Pearson’s correlation between the predicted chromatin accessibility and the real chromatin accessibility across different cell types for each locus [12].

Another challenge is the low sample number (~50) compared with high dimensional features (p > 30,000). over-fitting would be inevitable if we use normal linear regression. To solve the problem, we will firstly perform dimension reduction based on prior biological information: Enhancers normally regulate genes within 1Mb [14]. Thus, we will predict the chromatin accessibility using genes within 1Mb around the center of the DHSs. This will greatly decrease the number of features for prediction.

For model training, we will randomly select 80% of the DNase-seq and RNA-seq matched data set as training data, and use the left 20% (around 10) samples as test data. Then we will evaluate the performance of regularized linear regression (Lasso regression and Ridge regression), generalized additive model with smoothing spline (GAM), principal components regression and partial least squares regression by cross-validation and comparing cross-cell-type correlation. The model with the highest cross-cell-type correlation would be applied to test samples for performance evaluation.

**References**

1. Meisterernst M. Faculty of 1000 evaluation for Crystal structure of the nucleosome core particle at 2.8 A resolution [Internet]. F1000 - Post-publication peer review of the biomedical literature. 2001. doi:10.3410/f.1002097.20905

2. Radman-Livaja M, Rando OJ. Nucleosome positioning: how is it established, and why does it matter? Dev Biol. 2010;339: 258–266.

3. Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21: 456–464.

4. Henikoff S. Nucleosome destabilization in the epigenetic regulation of gene expression. Nat Rev Genet. 2008;9: 15–26.

5. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132: 311–322.

6. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 2006;16: 123–131.

7. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218.

8. Reeves R, Jones A. Genomic transcriptional activity and the structure of chromatin. Nature. 1976;260: 495–500.

9. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. Epigenetics Chromatin. 2014;7: 33.

10. Shu W, Chen H, Bo X, Wang S. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. Nucleic Acids Res. 2011;39: 7428–7443.

11. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30: 207–210.

12. Zhou W, Sherwood B, Ji Z, Du F, Bai J, Ji H. Genome-wide Prediction of DNase I Hypersensitivity Using Gene Expression [Internet]. 2016. doi:10.1101/035808

13. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: Roadmap for regulation. Nature. 2015;518: 314–316.

14. Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter–enhancer interactions and bioinformatics. Brief Bioinform. 2015; bbv097.