Association of enhancers and genes they regulate

Yunxiao Zhang
Guided by Sofia Kyriazopoulou
12/14/12
Final write-up

Overview

Enhancers are DNA elements that specifically activate the expression of their target genes. They are usually associated with modifications on the histones, the proteins that facilitate the organization of DNA in the cells, and thus can be identified based on such modifications.

Enhancers have features quite different from promoters, including high variability in nucleotide sequence, high flexibility in relative position to the target genes and highly complicated combinatorial regulation.

The enhancer sequences have a high diversity and account for the large degree of variability in gene expression between different cell types. The sequences can be far from genes and quite flexible in position. It's far from trivial to associate an enhancer sequence to its target genes.

On the other hand, the large numbers of enhancers and genes create a large set of combinations, and hence high correlations readily occur simply by chance, creating a host of false-positive results.

The project aims to associate enhancer elements to targets, based on the activity of enhancers and the expression level of genes. The activity of enhancers is variable in different cell types and accounts for different gene expression levels. The active enhancers are first recognized by epigenetic modifications. Then the enhancers in action are fed into a Generalized Linear Model (GLM) to predict the expression level of a host of genes. Then with parameter selection, the enhancers most relevant to the gene expression level are identified and the association is established.

Active enhancers are characterized by H3K27 acetylation peak whereas the transcription start sites (TSS) have enriched H3K4 methylation peaks (Fig 1). Some H3K27 acetylation peaks will be proximal to the H3K4 methylation and highly correlated with gene expression level. These proximal peaks tend to make other acetylation peaks trivial when fit into the GLM model. Hence only the subset of H3K27 acetylation peaks without overlapping H3K4 methylation peaks are recognized as active enhancers in the project.

Fig1. Epigenetic modifications of enhancers and transcription start site (TSS)

Genome is divided into functional partitions by insulator elements, so that long range action of enhancers is limited. Besides, significant association would be offset by background noise if too many peaks were included in analysis. Therefore, active enhancers are subsequently divided into subsets based on functional partition of the genome. The GLM model only considers the enhancers in the same functional partition as the gene and will be more sensitive to relevant enhancer elements.

In order to select the most relevant enhancer elements, the GLM model is fitted via penalized maximum likelihood with L1-regularization. Cross-validation determines the best lambda value for L1-regularization and selects the parameters. The model is then refit with the selected parameters. Peaks with coefficients that significantly deviate from 0 are considered as the enhancer in action.

In order to validate the results, False-Discovery Rate (FDR) is estimated. The genes are shuffled, so that the peaks are fitted to a random gene. The same criteria are used to pick significant associations. The associations found this way were considered as false-discoveries. The FDR is estimated by the following formula

$$\mathrm{FDR} = \frac{\text{Total nmber of false discovery}}{\text{True iscoveries} \times \text{Shuffle}}$$

The threshold for significance will then be tweaked to minimize FDR.

Raw Data

Chromatin-Immunoprecipitation (ChIP) and RNA-Sequencing (RNA-Seq) Data from lymphoblasts in 13 individuals were generated by the Snyder Lab at Stanford and preprocessed by Sofia Kyriazopoulou to obtain the signal at~~.~~ H3K4 methylation and~~,~~ H3K27 acetylation peaks as well as the expression values of genes. The data are also processed with DESeq to find differential expression patterns. Only genes that show differential expression in at least 10 pairs of the samples are included in the enhancer prediction.

Preprocessing of raw data

For the genes with very low expression, fold-change tends to be overestimated due to experimental noise. To reduce the effect of noise on low-signal genes/peaks, the signal was variance stabilized using the asinh function. The signal across all individuals was quantile normalized to account for experimental differences between the cell lines.

The peak data are also similarly processed, to get the mean counts from replicate experiments and then performed asinh quantile normalization.

Partition of Genome

The genome is partitioned with DESeq data. Genes with differential expression and the flanking regions without differential expression are defined as a functional partition. Partitions with too short lengths are merged.

Fig 2. Function partition of the genome

The real functional partition in genome is formed by insulator elements and the genes within a partition will tend to have similar trends in expression. The partitions obtained from DESeq data is an approximation of the real partitions. The length of partitions are well-controlled to be less than 1Mb, so the actual partitions obtained this way are non-overlapping with gaps. The regions without significant differential expression are depleted of partitions. This approach for partitioning is appropriate for finding associations in enhancers.

The mean and median of the length of partitions based on DESeq p-value 1e-4, 1e-3, 1e-2, 1e-1 is plotted as follows.

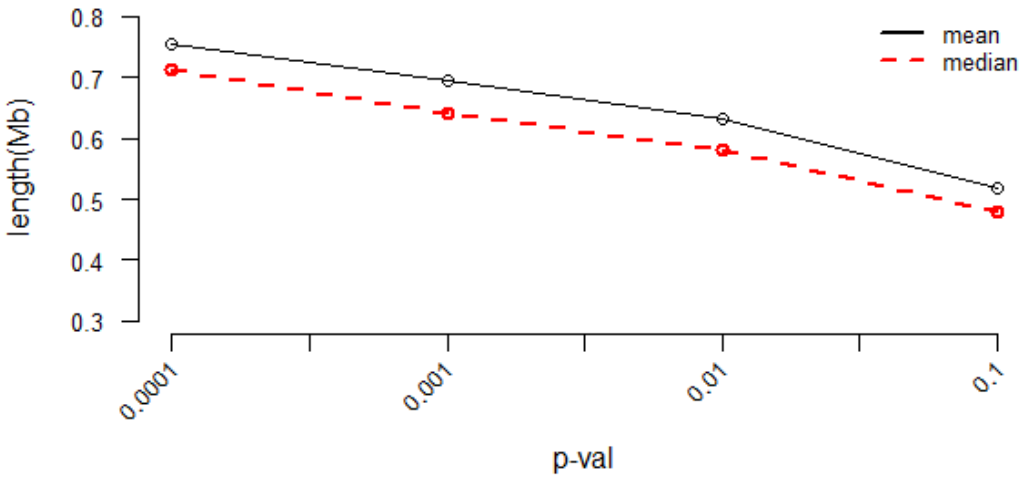

Fig 3. Change of length as p-value

The length of partitions decreases gradually with less stringent p-value.

Chromosome 1 is shown below to shown the distributions of the partitions in the genome..

Less stringent p-value yields a host of small partitions. The final p-value is chosen to ensure that the size of partitions are not too small and the partitions cover all the differentially expressed genes.

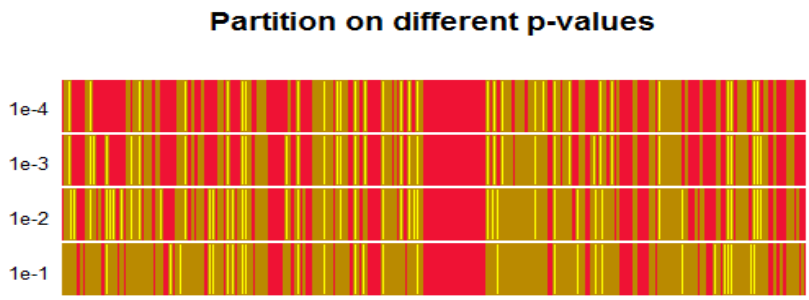

Fig 4. Partitions in the genome

L1-regularization

The GLM is fitted with L1 regularization. In this way, only the parameters (H3K27 acetylation peaks of enhancers) highly relevant to gene expression have non-zero coefficients. Cross-validation is used to determine the lambda value with the least deviation. Parameters (the peaks) with non-zero coefficients are chosen as the candidates. The selected parameters are then refit in GLM. The p-value for coefficient to be zero is calculated. Only parameters with p-value less than a threshold are considered significant and recorded as a hit.

The number of hits and estimated FDR is as follows.

Criteria	Number of Hits	FDR
p-value<0.01	224	0.284
p-value<0.001	69	0.282
p-value<0.01& log-likelihood>20	77	0.222
p-value<0.01& log-likelihood>30	40	0.115

Discussion

In previous studies[1,2], diverse cell types were used to include a major change in the transcriptomes, so that the differential expression would be common and associations would be abundant. In these studies, major enhancers for determinants of the cell types would be discovered.

In this project, the data are derived from lymphoblast cell lines from 13 individuals. As only one cell type is used, the differential expression will reflect minor differences between individuals and tend to be rare. The associations of enhancers with genes discovered in this data set will thus be rare but related with personal genomics. The nature of the discovery determines that most of the previously established associations will not be present in the results.

The algorithm for this project discovered a very tiny set of the enhancers. Many of them are not reported in previous studies. The FDR is still not satisfactory, but can be improved by finely-tuned cut-off parameters in the algorithm.

1 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232
nature11232 [pii] (2012).

2 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906
nature09906 [pii] (2011).