Snapshot: clustering and visualizing epigenetic history during cell differentiation

Guanjue Xiang, Belinda Giardine, Lin An, Chen Sun, Cheryl A. Keller, Elisabeth Heuston, David Bodine, Ross C Hardison, Yu Zhang

Abstract

Epigenetic modification of chromatin plays a pivotal role in regulating gene expression during cell differentiation. The scale and complexity of epigenetic data pose a significant challenge for biologists to identify the regulatory events controlling each stage of cell differentiation. Here, we present a model-free method, called Snapshot, that uses epigenetic data to generate a hierarchical visualization for the DNA regions segregating with respect to chromatin state along any given cell differentiation hierarchy of interest. Different cell type hierarchies may be used to highlight the epigenetic history specific to particular lineages of cell differentiation. We demonstrate the utility of Snapshot using data from the VISION project, an international project for VaIIdated Systematic IntegratiON of epigenomic data in mouse and human hematopoiesis.

Availability and implementation: https://github.com/guanjue/snapshot

1. Introduction

Multiple consortium projects have generated thousands of epigenomic datasets, and integration of such data has become a powerful way to understand the biological meaning of the combinations of epigenetic modifications(ENCODE Project Consortium, 2012; Bernstein *et al.*, 2010; Yue *et al.*, 2014). A commonly used method for studying epigenetic patterns in multiple cell types is to first perform peak calling on individual data, and then cluster the peaks according to the patterns of their signal intensity across all cell types(Corces *et al.*, 2016; Spencer *et al.*, 2015). These peaks are usually considered as candidate cis-regulatory elements (ccREs), and their epigenetic signal pattern can reflect their role in organismal development or cell differentiation. The groups of ccREs with active epigenetic marks are often the target of transcription regulatory machinery(Huang *et al.*, 2016). However, this approach has several limitations. Firstly, these methods do not take into account of any existing biological knowledge about the cell types. Secondly, interpreting the biological meaning and visualizing the identified ccRE clusters can be difficult, especially when the number of cell types is large. Here, we present a model-free method to cluster and visualize ccREs during cell differentiation. Our method takes into account known cell-to-cell relationships

in cell differentiation history, while providing the flexibility to analyze the ccREs along alternative histories. The method produces a comprehensive map of ccRE cluster patterns that can be used intuitively to compare, identify, and interpret the epigenetic history specific to each lineage of cell differentiation.

2. Methods

2.1 Clustering and visualizing ccREs based on their binary index

In contrast to unsupervised clustering methods, we developed a model-free method to cluster ccREs. While unsupervised clustering methods require pre-determining the number of clusters and can miss important clusters, our method can capture all distinct and recurring ccREs clusters. Each ccRE cluster portrays a distinct pattern of presence or absence of ccREs across the cell types examined. Specifically, we first perform peak calling on all cell types using an existing peak calling method, and we call the resulting peaks ccREs. Next, we use the binarized presence/absence status of ccREs at each location across all cell types to create a ccRE index to represent the unique combinatorial pattern of ccREs at the location. The number of bits in the index equals the number of cell types. The order of bits is the order of cell types derived from a user-provided cell differentiation tree. Each location with at least one ccREs across all cell types will receive an index. These indices readily classify the genomic locations into distinct ccRE clusters, since all ccREs with the same index are grouped into one cluster called an index-set. Each index-set contains a list of genomic locations that have the same ccRE presence/absence patterns across cell types. While each ccRE is in one index-set, it is of practical utility to restrict some further analyses to larger index-sets. Thus we filter out an index-set if its size is smaller than a user-specified threshold. Finally, we visualize the ccRE clusters in a heatmap. Each row in the heatmap is the ccRE pattern for each index-set, and each column is a cell type. The ccRE patterns are sorted by their indices in the heatmap. By our definition of the ccRE index, the ccRE patterns are separated if they have different ccRE status in a cell type; conversely they are clustered together if they have similar ccRE status in a cell type. This segregation is made initially at the top of the cell differentiation tree, and then it is repeated at each step along a user-specified cell lineage. Thus the major segrating ccRE clusters for a specific cell lineage can be well separated and illustrated in the heatmap. Furthermore, our method is flexible in that a user can specific a different lineage or order to the cell types, with different index-set maps focusing on distinctive lineages. See Figure 1a and results section for an example with more detailed explanation.

2.2 Visualizing epigenetic signal strength and their related functional state within each index-set.

To facilitate the interpretation of ccRE clusters, our data visualization package includes four sets of plots for each index-set: (1) a cell differentiation tree colored based on the intensity of epigenetic average signal in the index-set in each cell type (Figure 2a top); (2) violin plots of epigenetic signals in each cell type of each index-set (Figure 2a bottom). If the user provides a whole-genome functional annotation file, our package will further generate (3) an automatically colored cell differentiation tree based on the most frequent functional annotation in the index-set in each cell type (Figure 2b top); and (4) bar plots based on the proportion of each functional annotation in the index-set in each cell type (Figure 2b bottom). These different plots will together highlight the epigenetic activity across cell types and the associated functional annotations and their enrichments for each index-set, which enhances the interpretation of the functional roles of each ccRE cluster during cell differentiation.

2.3 Input and Main Options

We implement Snapshot as a python package with a graphical user interface. Snapshot takes the following files as input: (1) peak calling results from epigenetic data in different cell types in bed format (Kent *et al.*, 2002); (2) signal strength of each peak in the merged peak file in bed format; (3) functional annotation labels in bed format; (4) a list of colors for different functional annotations to be used in the heatmap; and (5) a list of files containing the input file names and the corresponding content labels in the output figures. The order of the input file names in peak file name list will be used as the cell type order in the index-set visualization. Snapshot uses bedtools(Quinlan, 2014) to handle most of the operations on the bed files. In terms of parameters, the user only needs to provide the minimum number of peaks that a biological meaningful index-set must contain.

3. Results

Here, we demonstrate our visualization package by analyzing the ATAC-seq data generated by the VISION project (**V**al**I**dated **S**ystematic **I**ntegrati**ON** of hematopoietic epigenomes) (Philipsen and Hardison, 2017; Oudelaar *et al.*, 2017). We first performed peak calling on the ATAC-seq data in 18 hematopoietic cell types using the peak calling software called Homer(Heinz *et al.*, 2010). The ATAC-seq data reveal genomic intervals that are accessible to nucleases, which is a feature associated with almost all cREs, and thus we treat the resulting peaks as ccREs. Next, we assigned each location with at least one ccREs an 18-digit index, where each digit corresponds to the presence (1) or absence (0) of the ATAC-seq peak in each of the hematopoietic cell type. In theory there are 2^{18} possible combinations of the 18-digit index, but our heatmap only plots index-sets containing more than 200 genomic regions (Figure 1a). The threshold of 200

is based on the distribution of the number of genomic regions covered by the index-sets (Figure 1c). One heatmap obtained through our data visualization package shows the ATAC-seq average signal strength (Figure 1a) and another shows the most frequent functional state (Figure 1b) in each cell type of the index-set.

From the heatmaps, specific ccRE clusters (index-sets) displaying ATAC-seq histories of interest to the user can be discovered. For example, we focused on index-set 150, an index-set containing 214 genomic regions, because of its ATAC-seq signal and functional annotation features suggest that these may correspond to ccREs involved in erythroid gene activation. Specifically, the accessibility (ATAC-seq signal) of this index-set gradually increased from the progenitor cells to the erythroblasts (Figure 2a), and its most frequent functional state became active enhancer state (orange color) upon entering the erythroid differentiation lineage (Figure 2b). The functional state annotation were generated by the IDEAS 2D genome segmentation method(Zhang *et al.*, 2016) using the epigenomic data in the VISION project. These observations suggested that these ccREs are critical for erythroid differentiation. As one test of this hypothesis, we examined the Gene Ontology (GO) terms of genes associated with these regions using the GREAT tool (McLean *et al.*, 2010). The results confirmed that the ccREs in index-set 150 were indeed significantly associated with erythroid differentiation (Figure 2c). Furthermore, the most significant enriched transcription factor binding site motifs (from MEME-ChIP, (Machanick and Bailey, 2011)) were those for the GATA transcription factor family. Two GATA factors, GATA1 and GATA2, are critically important for erythroid cell differentiation (Figure 2d)(Katsumura *et al.*, 2017).

4. Discussion

We developed the novel tool Snapshot that automatically generates cell differentiation associated heatmaps highlighting important ccRE clusters for lineage specific epigenetic events and their associated functions. It is a model-free clustering method that does not require a predetermined number of clusters and that can identify all abundant ccRE clusters. The results are more complete and more readily interpretable than those from conventional k-means and hierarchical clustering. Furthermore, its individual index-set data visualization module enables users to associate ccRE clusters with informative functional annotations, which can be genome segmentation results from epigenetic marks, ChIP-Seq data on protein-DNA interactions, and/or sequence information such as TF binding motifs. Finally, the graphical user interface and simple input format of Snapshot make it a handy tool for analyzing most genomic features across multiple cell types. Taken together, Snapshot can facilitate the discovery and interpretation of ccREs that are critical for lineage specific cell development. While we have described Snapshot in terms of its utility for analysis of ccREs across

differentiation, it can be applied to any progression of cell types, such as in response to hormones or signaling factors or along a developmental series.

ACKNOWLEDGEMENTS

The work is supported by NIH grants GM121613 and DK106766.

Reference

Bernstein,B.E. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.

Corces,M.R. *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, **48**, 1193–1203.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Heinz,S. *et al.* (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.

Huang,J. *et al.* (2016) Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Dev. Cell*, **36**, 9–23.

Katsumura,K.R. *et al.* (2017) The GATA factor revolution in hematology. *Blood*, **129**, 2092–2102.

Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Oudelaar,A.M. *et al.* (2017) Between form and function: the complexity of genome folding. *Hum. Mol. Genet.*, **26**, R208–R215.

Philipsen,S. and Hardison,R.C. (2017) Evolution of hemoglobin loci and their regulatory elements. *Blood Cells Mol. Dis.*

Quinlan,A.R. (2014) BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics*, **47**, 11.12.1–34.

Spencer,D.H. *et al.* (2015) Epigenomic analysis of the HOX gene loci reveals mechanisms that may control canonical expression patterns in AML and normal hematopoietic cells. *Leukemia*, **29**, 1279–1289.

Yue,F. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**, 355–364.

Zhang,Y. *et al.* (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.*, **44**, 6721–6731.

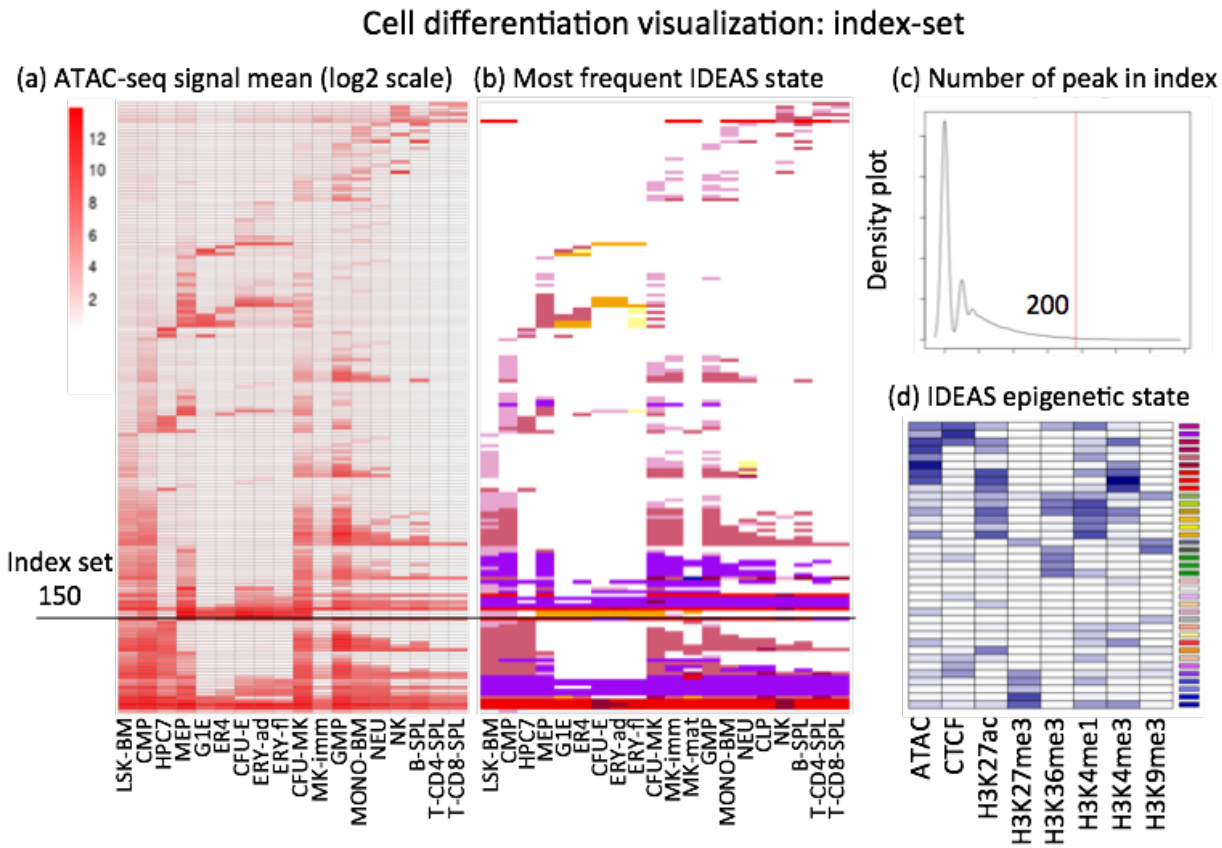

Figure 1: The heatmap of index-sets. (a) The heatmap of index-set colored by the average ATAC-seq signal in each cell type. (b) The heatmap of index-set colored by the most frequent functional annotation in each cell type. (c) The density plot of the number of genomic region covered by the index-set. (d) The color code and epigenetic composition of functional annotation used in (b).

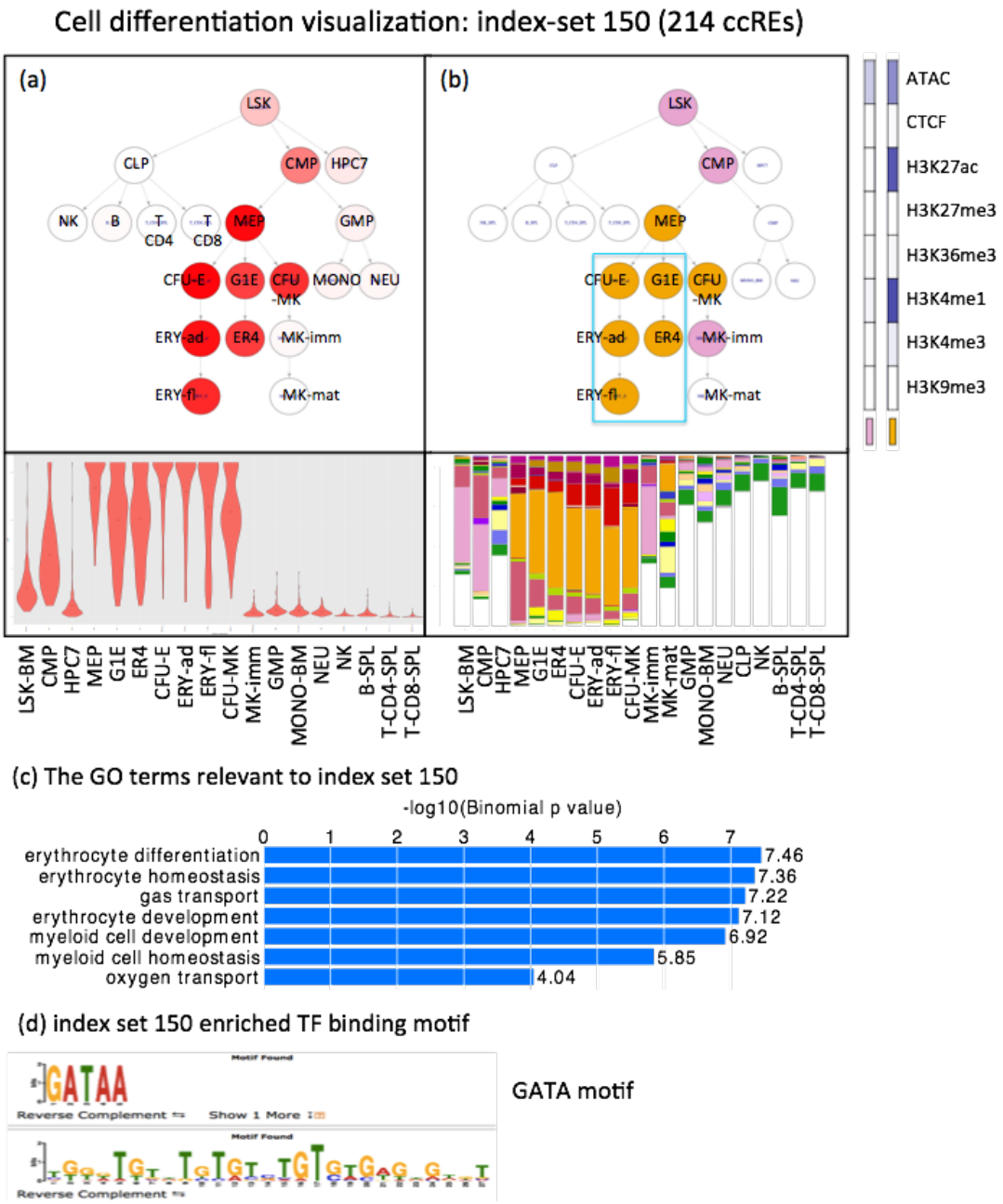

Figure 2: The data visualization for index-set-150 and corresponding GO analysis and MEME-ChIP TF binding motif analysis. (a) The hematopoietic cell differentiation tree colored by the average ATAC-seq signal in each cell type of the index-set-150. The violin plot represents the distribution of ATAC-seq signal in each cell type of the index-set-150 is in below. (b) The same cell differentiation tree colored by the most frequent functional annotation in each cell type of the index-set-150. The two most frequence functional annotation in erythroblasts lineage. The bar plot based on the proportion of each functional annotation in each cell type of the index-set-150 is below the cell differentiation tree. (c) The index-set-150 relevant GO term. (d) The index-set-150 significantly enriched TF binding motif from MEME-ChIP analysis.