Deciphering Epigenetic Regulation with Single Cell ATAC-Seq

Introduction

Understanding how cells differentiate during development is fundamental to the study of biology. Characterizing the regulatory mechanisms underlying these processes will provide valuable information in understanding not only normal development but also perturbed development, which can lead to developmental disorders and cancer in many cases.

The Chromium Single Cell ATAC Solution (scATAC-seq) is a powerful epigenomic tool to unravel cell type-specific regulatory networks. It has previously been demonstrated that this solution enables the profiling of accessible chromatin regions from tens of thousands of single cells at a time, thus enumerating active DNA regulatory elements, such as cis-acting DNA elements (e.g., enhancers and promoters) and trans-acting factors (e.g., transcription factors, TF) (1). These accessible chromatin regions can be used to distinguish cell types within complex populations and infer transcriptional regulatory mechanisms, making scATAC-seq data especially useful for studying gene regulation at single cell resolution.

Here, we present two proof-of-concept analyses of scATAC-seq data from human bone marrow mononuclear cells (BMMCs). In our first analysis, we inferred developmental trajectories during hematopoiesis and identified dynamic *cis*- and *trans*-regulatory elements in myeloid, lymphoid, and erythroid lineages. When profiling a heterogeneous population of cells in a developmental process, cells with similar developmental states are often close to one another in the numerical representation of gene space or other molecular features. As a result, we computationally inferred a developmental trajectory as the hidden structure within the cell population. Moreover, individual cells could be placed to the trajectory, forming a numerical sequence of cells which mimicked the temporal order of cells in the differentiation process. Such computationally inferred "time" is called pseudotime (2,3). This framework has two advantages. First, transient cell states can be analyzed without the need to purify individual cell types or states, which may be difficult or impossible experimentally. Second, pseudotime analysis allows detection of differentially regulated features over the course of the trajectory, leading to deeper insights into key transcriptional programs and regulatory networks. Previous studies have demonstrated that computationally inferred pseudotime has a high level of agreement with the experimentally observed biological time of the developmental process (4,5).

Highlights

We demonstrate strategies for analyzing Single Cell ATAC-seq data to gain insights into the epigenetic regulation of gene expression profiles:

- Infer developmental trajectories for three lineages in the human hematopoietic system
- Identify differentially accessible *cis*- and *trans*-regulatory elements in developmental trajectories
- Construct regulatory networks of transcription factors
- Detect lineage-specific transcription factor interactions

In the second analysis, we constructed interaction networks of a set of key TFs in hematopoietic stem cells (HSC) and terminally differentiated erythroid and B cells. Traditionally, such networks have been generated by targeting one cell type and one TF at a time, an approach that can be challenging due to limited availability of purified cell types, complex experimental procedures, and a lack of knowledge of all relevant transcription factors in a developmental process. By contrast, scATAC-seq data enable the detection of a large number of TF regulation signatures within individual cell types, all in one assay.

In both analyses, we confirmed many known regulatory mechanisms of the human hematopoietic system, and we envision that the Chromium Single Cell ATAC Solution and analysis strategies will be further applied to better understand epigenetic regulation in diverse biological systems.

10x GENOMICS

Methods

Single Cell ATAC-seq Profiling of Human Bone Marrow Tissues

BMMCs and CD34+ cells were purchased from AllCells and processed according to the 10x Genomics Demonstrated Protocol-Nuclei Isolation for Single Cell ATAC Sequencing (Document CG000169). Libraries were prepared following the Chromium Single Cell ATAC Reagent Kits User Guide (Document CG000168; the Next GEM reagents may also be used for these experiments, Document CG000209) and sequenced to a depth between 20,000 and 50,000 raw read pairs per cell. Sequencing data was processed through the Cell Ranger ATAC pipeline v1.1.0, and *cellranger-atac aggr* was used to aggregate the BMMCs and CD34+ data. CD34+ cells were added due to the low proportion of CD34+ cells in healthy BMMCs.

Single Cell RNA-Seq Profiling of Human BMMCs

The same batch of BMMCs from AllCells was used to generate scRNA-seq data of ~6500 cells following the Chromium Single Cell 3' Reagent Kits User Guide (v3 Chemistry), (Document CG000183; the Next GEM reagents may also be used for these experiments, Document CG000204) ~40,000 raw reads pairs per cell were sequenced with 1500 median genes/cell and ~4300 median UMI counts/cell detected.

Single Cell Trajectory Analysis

For the subsequent analyses outlined in the Methods section, we employed third party tools not part of Cell Ranger ATAC pipeline v1.1.0. The results of these analyses are shown for demonstration purposes and are not supported by 10x Genomics.

We utilized UMAP for visualization due to several factors: 1) UMAP preserves as much local and more global distance in comparison to tSNE, 2) it is widely adopted by the single cell genomics community, and 3) it is scalable to datasets containing large numbers of cells. Latent semantic analysis (LSA) components 2 to 15 from the Cell Ranger ATAC pipeline output were used as input for UMAP. Removing the first principal component and potential multiplet clusters improved the robustness and the quality of the UMAP projection.

We used Monocle (version 3 alpha) for trajectory inference, with custom parameter settings for "min_dist=0.3", "n_neighbors=30", and "repulsion_strength=2". We chose the minimal spanning tree as the base graph building method due to its scalability and wide adoption. Monocle's "learnGraph" function and "simplePPT" algorithm were used to learn a minimal expansion tree. To find individual paths, we defined the node with maximum density of HSCs and multipotent progenitor (MPP) cells as the root node. We defined leaf node as the remaining nodes with only one connected node, and a path as the shortest path from a root node to a leaf node. The path-finding step was helpful in visualizing tree-based trajectory and simplified the pseudotime ordering process.

Gene Activity (GA) and Transcription Factor Deviation Score Calculation

We calculated gene activity scores using the R package Cicero (6), obtaining the peak-to-gene annotation and tSNE coordinates (as the reduced_coordinates) directly from the Cell Ranger ATAC output. We normalized the GA as log2(GA*1e6 +1) to make the score more interpretable (1,6).

To measure global TF activity with chromVAR (7), we obtained the input count matrix from the TF-barcode matrix of the Cell Ranger ATAC pipeline and selected the JASPAR motif database as the input motif database. TF deviation scores were then calculated following the recommended workflow of chromVAR (7).

Differential Expression Along Single Cell Developmental Trajectories

To define differentially accessible (expressed) regulatory elements and TFs (referred to as "genes" in general), we applied SpatialDE (8) on the gene activity scores and TF deviation scores against the pseudotime axis. SpatialDE decomposes gene expression variability into spatial and non-spatial components such that the significance of spatially variable genes have a high proportion of variance explained by the spatial component (8). For differential activity analysis, gene activity scores over pseudotime were used as SpatialDE's input. We chose a linear kernel to identify genes that are turned on or off during a specific stage of the trajectory. Genes gradually turned on or off during the trajectory were also identified.

Analysis of Transcription Factor Networks

A generic TF (TF1) was defined to regulate another TF (TF2) if TF1 had a putative binding site near the TF2 gene. TF1 putative binding sites were defined as peaks (accessible regions) containing a TF1 binding motif. The map between peaks and TF motifs was obtained from peak_motif_hits.bed (which can be found in the _PEAK_ANNOTATOR/SCAN_MOTIFs directory of a Cell Ranger ATAC pipeline run). A motif was defined as "near" a TF2 gene when its start coordinate was within 5 kb of the transcription start site of the TF2 gene (either up or downstream).

To evaluate whether or not TF1 regulates TF2, we calculated a score which combined the accessibility of the peak (TF1 motif near TF2 gene) and the gene activity score of TF1. The "expression" of the peak was normalized cut sites count at the peak.

$$\text{Score}_{\text{cell\_i}} = \text{peak\_exp}_{\text{cell\_i}} * \text{GA}_{\text{cell\_i}}$$
$$\text{Score} = \text{mean}(\text{Scores}_{\text{across\_all\_cells}})$$

To identify cell type-specific TF interactions, we used the peak expression and gene activity scores of every cell of a specific cell type. Intuitively, TF1 regulation on TF2 would have a relatively high score in cell type A when the TF1 binding motif near the TF2 gene was highly accessible and when TF1 was highly active in cell type A. We used visNetwork R package to visualize networks in HSCs, erythrocytes, and B cells.

Results

Construction of Developmental Trajectories in Human BMMCs

To demonstrate the power of scATAC-seq in trajectory analysis, we used scATAC-seq data from 10,321 BMMCs and 9,084 CD34+ cells, which collectively included stem cells and cell types spanning the myeloid, erythroid, and lymphoid lineages (Figure 1A). We identified 19 major cell types based on enrichment of cell type-specific peaks (Figure 1B) (For details on cell type annotation, refer to CG000234. The CD34+ progenitor population included hematopoietic stem cells (HSC), multipotent progenitor cells (MPP), common myeloid progenitors (CMP), granulocyte-macrophage progenitors (GMP), lympho-myeloid primed progenitors (LMPP), common lymphoid progenitors (CLP), and megakaryocyte-erythroid progenitors (MEP). Terminally differentiated cell populations included B cells, erythrocytes, CD4+ T cells, CD8+ T cells, and natural killer (NK) cells. We expected a very small proportion of granulocytes (Gn) in our BMMCs because of the enrichment of mononuclear cells in sample preparation. 19 cell clusters were identified and grouped into two larger, well-separated partitions (Monocle, Methods), partition 1 containing the progenitor cells and partition 2 containing the more terminally differentiated T and NK cells (Fig 1C). We focused on the progenitor cell-containing partition 1 for the developmental trajectory analysis.

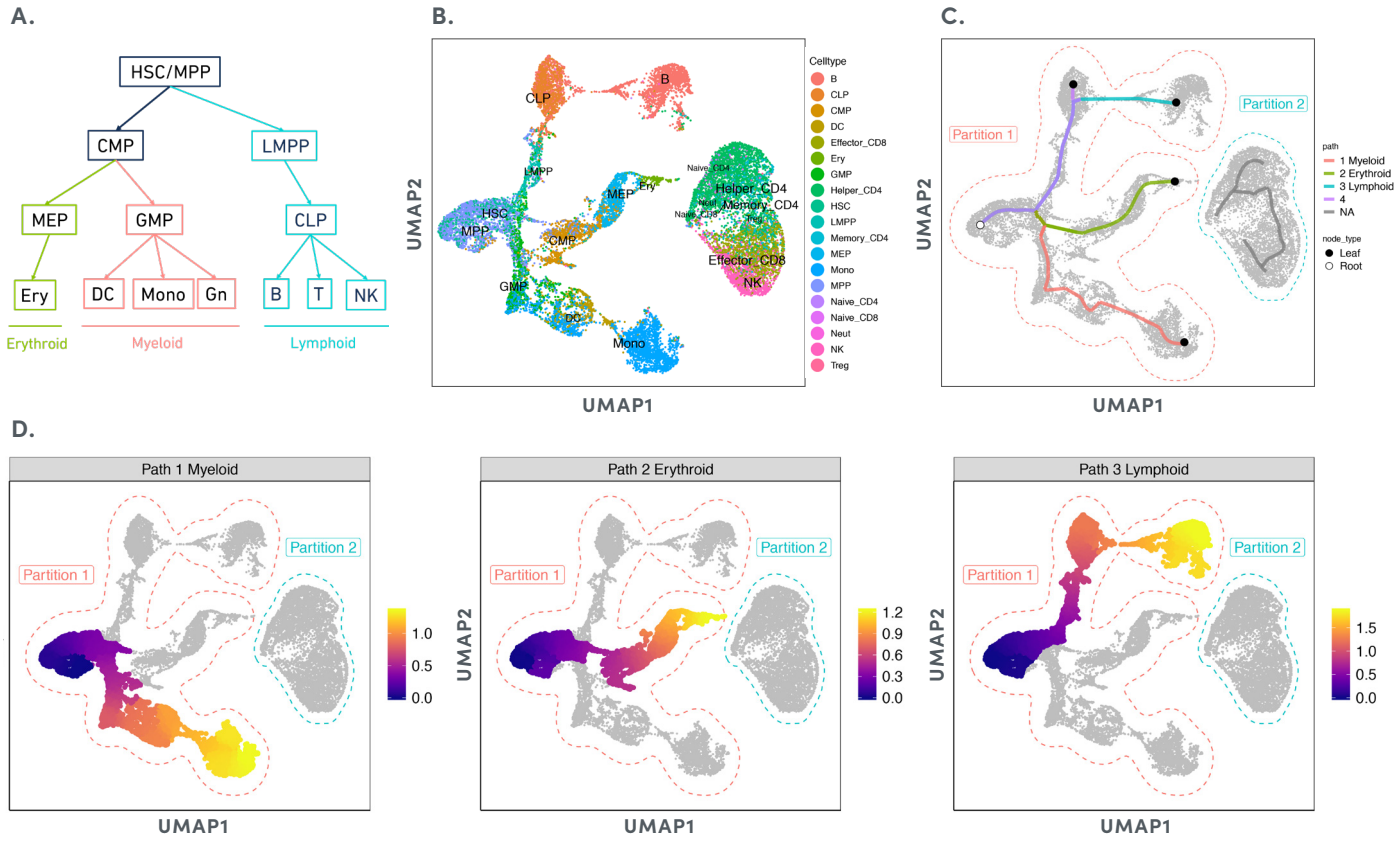

Figure 1 Inferred developmental trajectories in human BMMCs. A. Cell type hierarchy of hematopoiesis. B. UMAP projection of human bone marrow mononuclear cells. C. Cell partition and inferred trajectory paths for cell partition group 1. Among the four paths identified, paths 1, 2, and 3 roughly represent development lineage of myeloid, erythroid, and lymphoid cells. Path 4 is highly redundant with path 3 and stops at a progenitor cell stage. Both root (empty circle) and leaf node (solid circle) types are labeled. There is no root node in the T cell partition in which all paths will be excluded from the downstream analysis. D. Pseudotime distribution for each lineage trajectory. HSC: Hematopoietic stem cells; MPP: Multipotent progenitor cells; CMP: Common myeloid progenitors; CLP: Common lymphoid progenitors; GMP: Granulocyte-macrophage progenitors; Mono: Monocytes; DC: Dendritic cells; Gn: Granulocytes; LMPP: Lympho-myeloid primed progenitors; MEP: Megakaryocyte-erythroid progenitors; Ery: Erythrocytes; B: B cells; T: T cells; NK: Natural killer cells.

Using Monocle, we identified four developmental paths originating from the HSC root node in partition 1 (Methods). Path 4 was excluded from further analysis as it largely overlaps with path 3. Based on annotations of cells along each path, we assigned paths 1, 2, and 3 to myeloid, erythroid, and lymphoid lineages, respectively (Figure 1C). We then separated cells by path and computed pseudotime for each path (Figure 1D).

The myeloid lineage originates from MPPs and HSCs, which give rise to CMPs and GMPs, ultimately maturing to dendritic cells and monocytes (Figure 2). The erythroid lineage also starts with MPPs and HSCs, however these cells differentiate to CMPs and MEPs, later becoming terminally differentiated erythrocytes. Interestingly, 14.5% of the GMPs detected were located in the lymphoid lineage, the majority being in the myeloid lineage. This is likely due to ambiguity in annotation of GMPs. Overall, the distribution of major cell types along each path largely recapitulates known hematopoietic differentiation trajectories, demonstrating that scATAC-seq data can be used to accurately infer developmental trajectories (Figure 2).

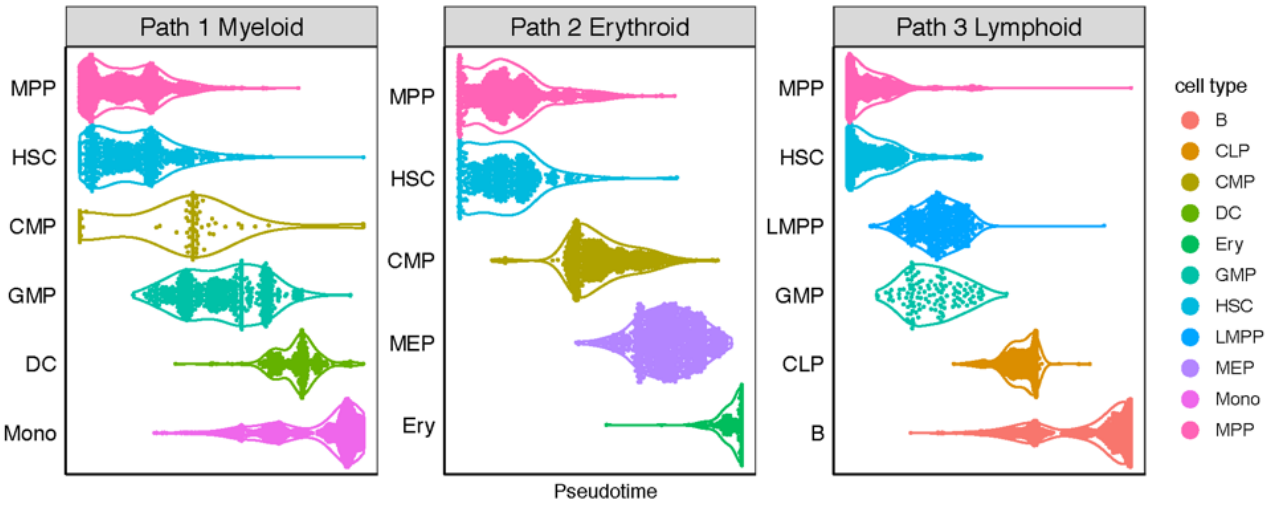

Figure 2 Dynamic accessibility profiles of *cis*- and *trans*-regulatory elements in hematopoiesis. Major cell type density distribution in pseudotime in each lineage trajectory. Each dot inside the density panel represents a single cell. A small population of CMPs (3% of total) were placed at the beginning of the pseudotime in the myeloid path, likely due to ambiguity of cell type annotation and batch effect from data aggregation.

Analysis of Chromatin Accessibility Dynamics in Developmental Trajectories

The trajectory graph laid a foundation for pseudotime ordering, a dimension transformation procedure that converts the two spatial dimensions of the UMAP projection to a single continuous time dimension. Pseudotime ordering allows detection of differentially regulated genomic features.

We first visualized a few representative transcription factors for activity patterns along the pseudotime axis. For example, in B cell development, HOXA9 is gradually turned off from stem cell to committed progenitors. EBF1 is activated in CLPs and pre/pro-B cells, prior to maturation of the B cells, at which point SOX2 (POU2f2) and IRF4 are activated (Figure 3). All observations from the pseudotime analysis are consistent with the known regulatory pathways in human hematopoietic differentiation.

Next, we sought to determine the differentially accessible *cis*- and *trans*-regulatory elements genome-wide in each of the developmental trajectories. We identified significantly variable TFs (ChromVar TF deviation scores) that are turned on in early, intermediate, or late stages of each lineage trajectory (Figure 4A). For example, HOXA9 motif sites are highly accessible in early stem cell populations, displaying a similar pattern to that of the *cis*-regulatory enhancers of HOXA9. The GATA family (GATA1-6) of transcription factors show increased accessibility in committed erythroid progenitors (megakaryocyte–erythroid progenitor, MEPs) but not in the myeloid or lymphoid lineages. On the other hand, the C/EBP family of transcription factors (CEBPA, CEBPB) is turned on in committed myeloid progenitors but not in the erythroid lineage. Lymphocyte-specific transcription factors, such as SPI1, ETS1, TBX21, and EOMES, all showed increased TF activity, i.e. accessibility of the motif sites, in the lymphoid lineage.

We also took advantage of the SpatialDE algorithm (8) and identified genes whose *cis*-regulatory elements (gene activity scores) have significant variation along the trajectory pseudotime, including genes with increased or decreased activity along the trajectory, as well as transitional genes that are turned on at specific stages of the pseudotime trajectory (Figure 4B). For example, HOXA9 gene activity is enriched early in HSCs and MPPs. Uroporphyrinogen III synthase (UROS), a key enzyme in the heme biosynthesis pathway, is exclusively activated in late stages of the erythroid lineage trajectory, while B-cell marker MS4A1 is exclusively expressed in B cells.

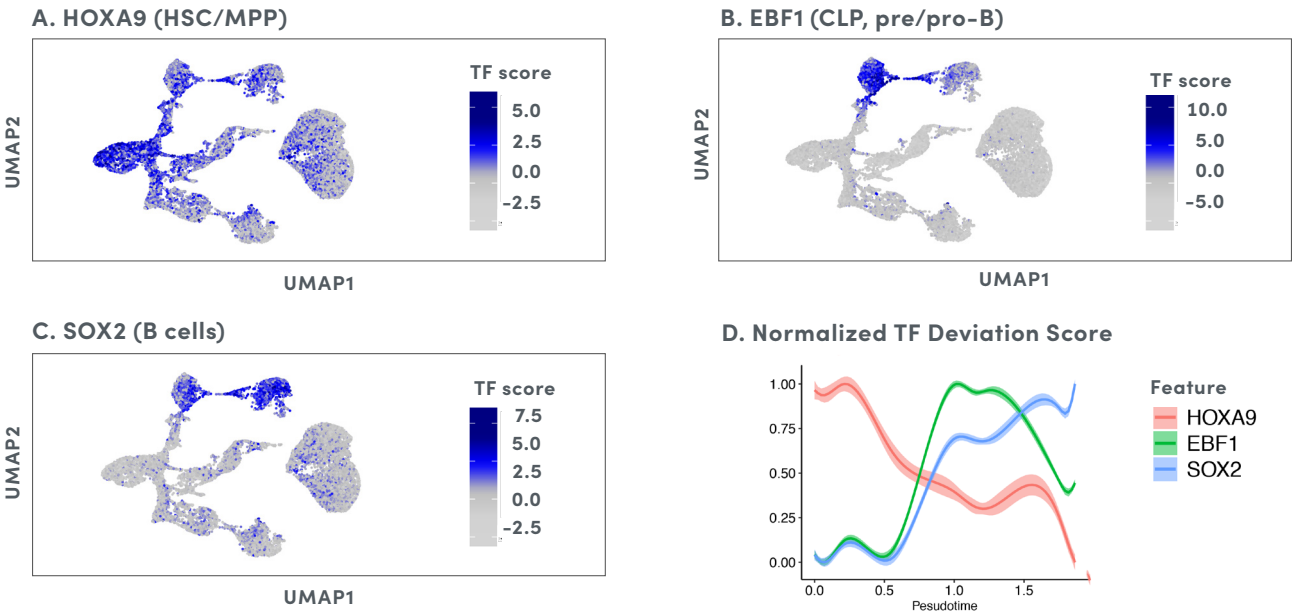

Figure 3 Transcription factor dynamics in lymphoid lineage trajectory. A. HOXA9 activity is enriched in HSCs and MPPs. B. EBF1 activity is enriched in CLP and B-cell progenitors. C. SOX2 activity is enriched in B cells. D. Cell type-specific transcription factors HOXA9, EBF1, and SOX2 are activated in early, intermediate, and late stages of lymphoid lineage trajectory, respectively. Curves shown are smooth splines fitted to ChromVAR deviation scores normalized by min-max in pseudotime dimension.

Detection of Cell Type-Specific Transcription Factor Regulatory Networks

Transcription factors are key drivers of developmental processes. It is well established that TFs regulate each other to initiate and maintain transcriptional programs (9–14). Better understanding of TF interaction networks can lead to deeper insight into complex regulatory mechanisms. To demonstrate the use of scATAC-seq data in the construction of TF regulatory networks, we used the aggregated scATAC-seq data from BMMCs and CD34+ cells (Figures 1–4) to infer well-characterized TF regulatory networks in hematopoiesis. We focused on nine TFs in HSCs, erythrocytes, and B cells with previously characterized auto- and cross-regulatory networks, namely GATA1, GATA2, TAL1, SPI1, ELF1, HES1, MYB, PAX5, and EBF1 (9,10,15). We successfully detected most of the expected interactions, some of which are known to be present in all three cell types, and others that are enriched in specific cell types only. For example, SPI1 regulation on MYB is detected in all 3 cell types (Figure 5A). In contrast, GATA2 and TAL1 regulation is preferentially detected in HSCs, and GATA1 autoregulation is preferentially detected in erythrocytes (Figure 5A).

The activity of a transcription factor gene and the accessibility of its target gene binding site are both important contributors to cell type-specific networks. In order for TF1 to regulate TF2 in a specific cell type, the TF1 gene needs to be active in that cell type and the TF1 binding site at the TF2 gene needs to be accessible. For example, GATA1 autoregulation is specific in erythrocytes because GATA1 gene expression is active in erythrocytes (Figure 5C) and the GATA1 binding site is only accessible in erythrocytes (Figure 5B).

While scRNA-seq can provide cell type-specific gene expression in complex cell populations, its application to understanding TF regulatory networks can be greatly facilitated when combined with scATAC-seq data. First, gene activity scores from scATAC-seq can complement gene expression from scRNA-seq. For example, very few HES1 transcripts were detected in BMMCs by scRNA-seq (Figure 5C). By contrast, HES1 gene activity was detected in most cell types by scATAC-seq (Figure 5C), which is consistent with the activity of HES1 in blood cells (GTEx and ProteomicsDB, (15)). Second, the accessibility of putative binding sites from scATAC-seq improves the specificity of regulatory networks inferred from scRNA-seq data. For example, SPI1 regulation on TAL1 is mostly enriched in HSCs (Figure 5A) (15), although SPI1 and TAL1 expression is detected in both HSCs and erythrocytes (Figure 5C). The regulation of EBF1 by PAX5 and SPI1 also demonstrates the complementarity of gene expression specificity and TF binding specificity. The regulation specificity of PAX5 at EBF1 promoter can be explained by the dominant B cell-specific expression (Figure 5C). On the other hand, EBF1 regulation by SPI1 is mostly driven by B cell-specific binding events at the distal site 16 kb downstream of the EBF1 promoter (Figure 5B), while SPI1 expression is detected in multiple cell types (Figure 5C).

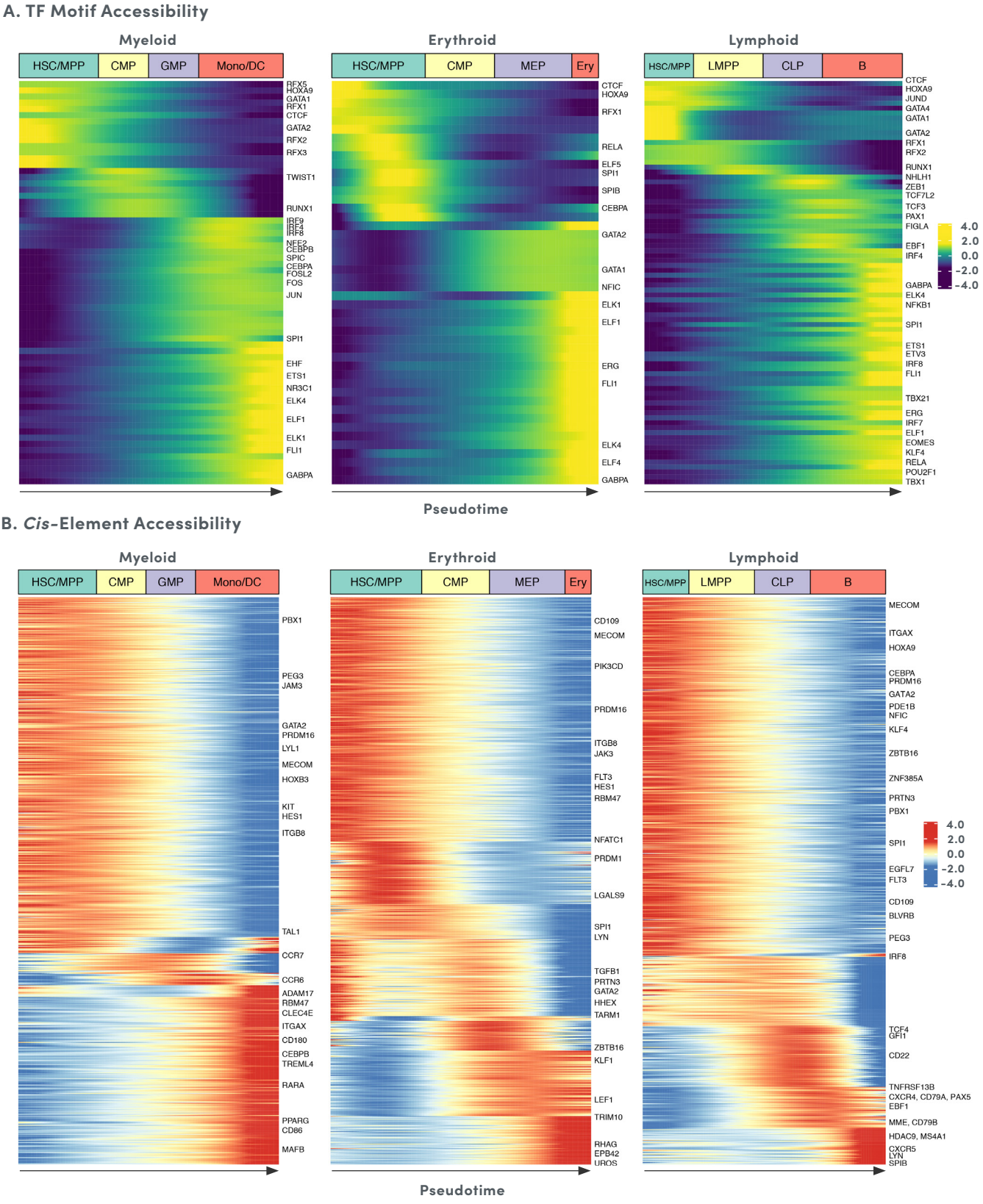

Figure 4 Dynamic accessibility profile of *cis*- and *trans*-regulatory elements in hematopoiesis. A. Global dynamics of top variable transcription factors in myeloid, erythroid, and lymphoid lineage trajectories identified by chromVAR. B. Top differentially accessible genes in myeloid, erythroid, and lymphoid lineage trajectories identified by SpatialDE.

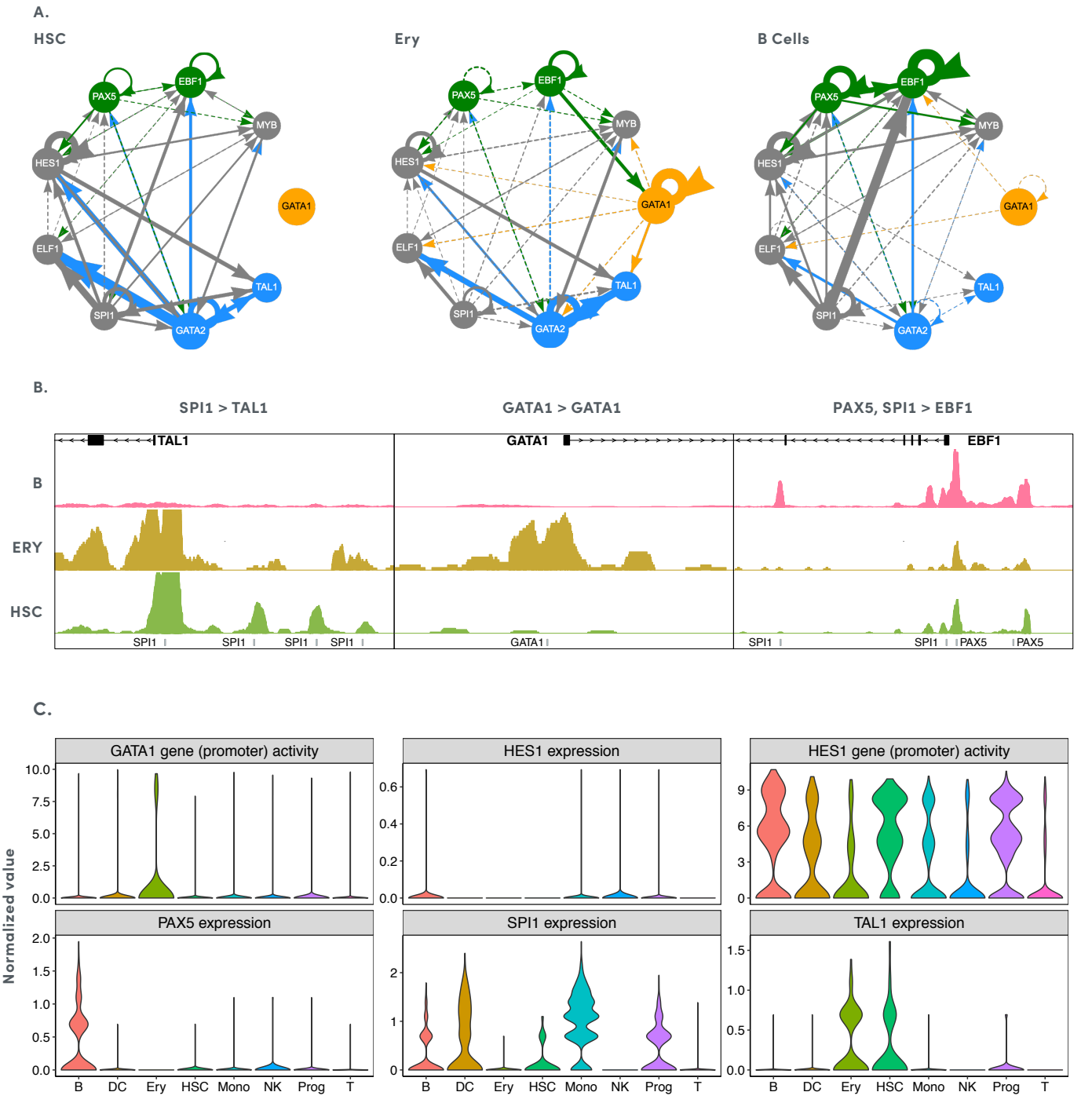

Figure 5 Cell type-specific transcription factor interaction networks. A. TF networks detected in HSCs (left panel), erythrocytes (middle panel), and B cells (right panel), focusing on the nine transcription factors shown. TFs are color-coded based on their known activity in hematopoiesis; blue represents HSC-specific, orange erythrocyte-specific, green B cell-specific, and grey represents a lack of cell type specificity. A solid arrow head from TF1 to TF2 represents a significant regulation of TF1 on TF2. A dotted arrow head from TF1 to TF2 represents a detected but non-significant regulation of TF1 on TF2. The width of the line is proportional to the regulatory score; the thicker the line, the more enriched the regulation in that cell type. B. Normalized accessibility of TF binding sites for PAX5, SPI1, and GATA1 at target genes EBF1, TAL1 and GATA1 locus in HSCs, Erythrocytes, and B cells. Position of TF binding site motifs are shown as grey boxes in the bottom row. Height of tracks are normalized to number of cells (GATA1 and EBF1 loci) and scaled to highlight differentially accessible peaks (TAL1 locus). C. Normalized gene expression and gene (promoter) activity for selected TFs in bone marrow cell types. Prog: hematopoietic progenitor cells, including all CD34+ cells in BMMC except the hematopoietic stem cells (HSCs).

10x GENOMICS

Conclusions

We demonstrate computational analysis strategies that use single cell ATAC-seq data to gain insights into epigenetic regulation, including constructing developmental trajectories and TF interaction networks within the human hematopoietic system. As these analyses cover a fraction of applications for single cell ATAC-seq data, researchers in the future will continue to use the Chromium Single Cell ATAC Solution, expanding its applications to learn more about diverse biological systems.

Resources

Datasets	go.10xgenomics.com/scATAC/datasets
Seminars	go.10xgenomics.com/scATAC/seminars
Application Notes	go.10xgenomics.com/scATAC/app-notes
Technical Support	go.10xgenomics.com/scATAC/support
Publications	go.10xgenomics.com/scATAC/pubs

Support

support@10xgenomics.com
10x Genomics
6230 Stoneridge Mall Road
Pleasanton, CA 94588-3260

References

1. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, *et al.* **Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion.** *Nat Biotechnol.* 37, 925-936 (2019).
2. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, *et al.* **The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.** *Nat Biotechnol.* 32, 381-386 (2014).
3. Saelens W, Cannoodt R, Todorov H, Saeys Y. **A comparison of single-cell trajectory inference methods.** *Nat Biotechnol.* 37, 547-554 (2019).
4. Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, *et al.* **A single-cell molecular map of mouse gastrulation and early organogenesis.** *Nature.* 566, 490-495 (2019).
5. Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. **Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis.** *Science.* 360 (2018).
6. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, *et al.* **Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data.** *Mol Cell.* 71, 858-871.e8 (2018).
7. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. **chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data.** *Nat Methods.* 14, 975-978 (2017).
8. Svensson V, Teichmann SA, Stegle O. **SpatialDE: identification of spatially variable genes.** *Nat Methods.* 15, 343-346 (2018).
9. Nutt SL, Kee BL. **The Transcriptional Regulation of B Cell Lineage Commitment.** *Immunity.* 27, 361 (2007).
10. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. **Circuitry and dynamics of human transcription factor regulatory networks.** *Cell.* 150, 1274-1286 (2012).
11. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, et al. **Core transcriptional regulatory circuitry in human embryonic stem cells.** *Cell.* 122, 947-956 (2005).
12. Kim J, Chu J, Shen X, Wang J, Orkin SH. **An extended transcriptional network for pluripotency of embryonic stem cells.** *Cell.* 132, 1049-1061 (2008).
13. Davidson EH. **A Genomic Regulatory Network for Development.** *Science.* 295, 1669-1678 (2002).
14. Yun K, Wold B. **Skeletal muscle determination and differentiation: story of a core regulatory network and its context.** *Curr Opin Cell Biol.* 8, 877-889 (1996).
15. Swiers G, Patient R, Loose M. **Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification.** *Dev Biol.* 294, 525-540 (2006).

Legal Notices

For 10x Genomics legal notices visit:
10xgenomics.com/legal-notices