TECHNICAL NOTE

Cell Type Annotation Strategies for Single Cell ATAC-Seq Data

Introduction

The starting point for interpretation of any single cell sequencing data is the annotation of cell clusters in a given dataset. Cell type annotation in single cell ATAC-seq data is challenging due to lack of specifically designed tools and use of unintuitive cis- and trans-regulatory elements in single cell ATAC-seq data. This Technical Note explores and demonstrates three different strategies that vary in the amount of bioinformatics expertise required for annotating cell types in single cell ATAC-seq data.

Methods

Human bone marrow mononuclear cells (BMMCs) and Fluorescence Activated Cell Sorting (FACS) enriched CD34$^+$ hematopoetic progenitor cells (AllCells) were processed according to the 10x Genomics Demonstrated Protocol - Nuclei Isolation for Single Cell ATAC Sequencing (Document CG000169). Single Cell ATAC libraries were prepared following the Chromium Single Cell ATAC Reagent Kits User Guide (Document CG000168) and sequenced at 20,000-50,000 raw read pairs per cell. The sequencing data were processed through the cellranger-atac count (v1.1.0) pipeline and the cellranger-atac aggr pipeline was used to aggregate BMMCs and CD34$^+$ cells data. A similar strategy can be implemented using the Chromium Single Cell ATAC v2 Reagent Kits User Guide (Document CG000496) and cellranger-atac count (v1.1.0 or later) pipelines.

The cell type annotation strategies outlined below are possible methods of cell type annotation in single cell ATAC-seq data and are not a part of the Cell Ranger ATAC Software or supported by 10x Genomics.

Strategy 1. Annotation Using Cis-Regulatory Elements

Single cell ATAC-seq data from 10,321 BMMCs and 9,084 CD34$^+$ cells were analyzed in Loupe Cell Browser 3.1.1. CD34$^+$ progenitors, CD4$^+$ T cells, CD8$^+$/NK cells, B cells, and monocytes/dendritic cells were labeled by visualization of promoter accessibility patterns for cell type marker genes (Figure 1). Cell type-specific cut site distribution was exported from Loupe Cell Browser by loading the fragments.tsv.gz to peak viewer and exporting cut sites per cell type per window.

Strategy 2. Annotation Using Cell Type-Specific Feature Set

A second method of cell type annotation employs a user-defined set of molecular features including cell type-specific peaks, gene activation scores of cell type markers, or motif accessibility of transcription factors with known regulatory roles. For example, to annotate cell types using cell type-specific peaks, a scoring scheme that computes the enrichment of cell type-specific peaks over background accessibility levels was applied to the single cell ATAC-seq data from 10,321 BMMCs and 9,084 CD34$^+$ cells. A unified set of 1.3 million peaks was curated by Epinomics from 29 FACS-sorted immune cell types to define the ATAC profiles of those cell types, based on previously published data (1). Cell type-specific peaks were defined as the top 200 enriched peaks of the selected cell type over all other cell types. Background was defined as 500 sets of 200 randomly selected peaks. The cell type generating the maximum enrichment score was annotated to the cell (Figure 2).

Strategy 3. Annotation Using RNA Sequencing Data as Reference

To annotate cell types using RNA-seq data, single cell ATAC-seq data were generated from embryonic and adult mouse brain tissues (see below) using the Chromium Single Cell ATAC Solution. The reference RNA-seq datasets for the embryonic and adult mouse brain tissue were derived from a previous study (2,3). Seurat v3.0 and Signac packages (4) were used to co-embed single cell ATAC-seq data and single cell RNA-seq data into a shared reduced dimension and predict cell types for ATAC-seq data based on distances to the pre-annotated cells in RNA-seq data (Figure 3).

Query (single cell ATAC)

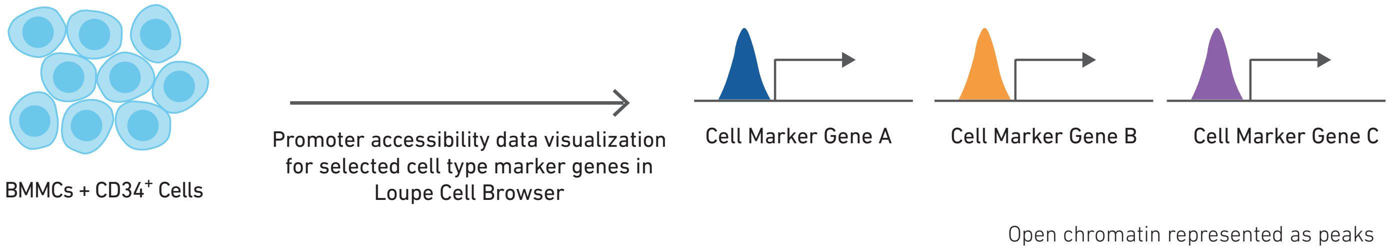
- P50 adult mouse cortex (3,927 cells)
- E18.5 mouse cortex, hippocampus & ventricular zone (4,115 cells)

Reference (single cell RNA)

- P30-40 mouse primary visual cortex & anterior lateral motor cortex (21,814 cells)
- P0 mouse cortex (7,614 cells)

Strategy 1: Annotation Using Cell Type-Specific Cis-Regulatory Elements

Methods


BMMCs + CD34⁺ Cells
Promoter accessibility data visualization for selected cell type marker genes in Loupe Cell Browser
Cell Marker Gene A
Cell Marker Gene B
Cell Marker Gene C
Open chromatin represented as peaks

Results

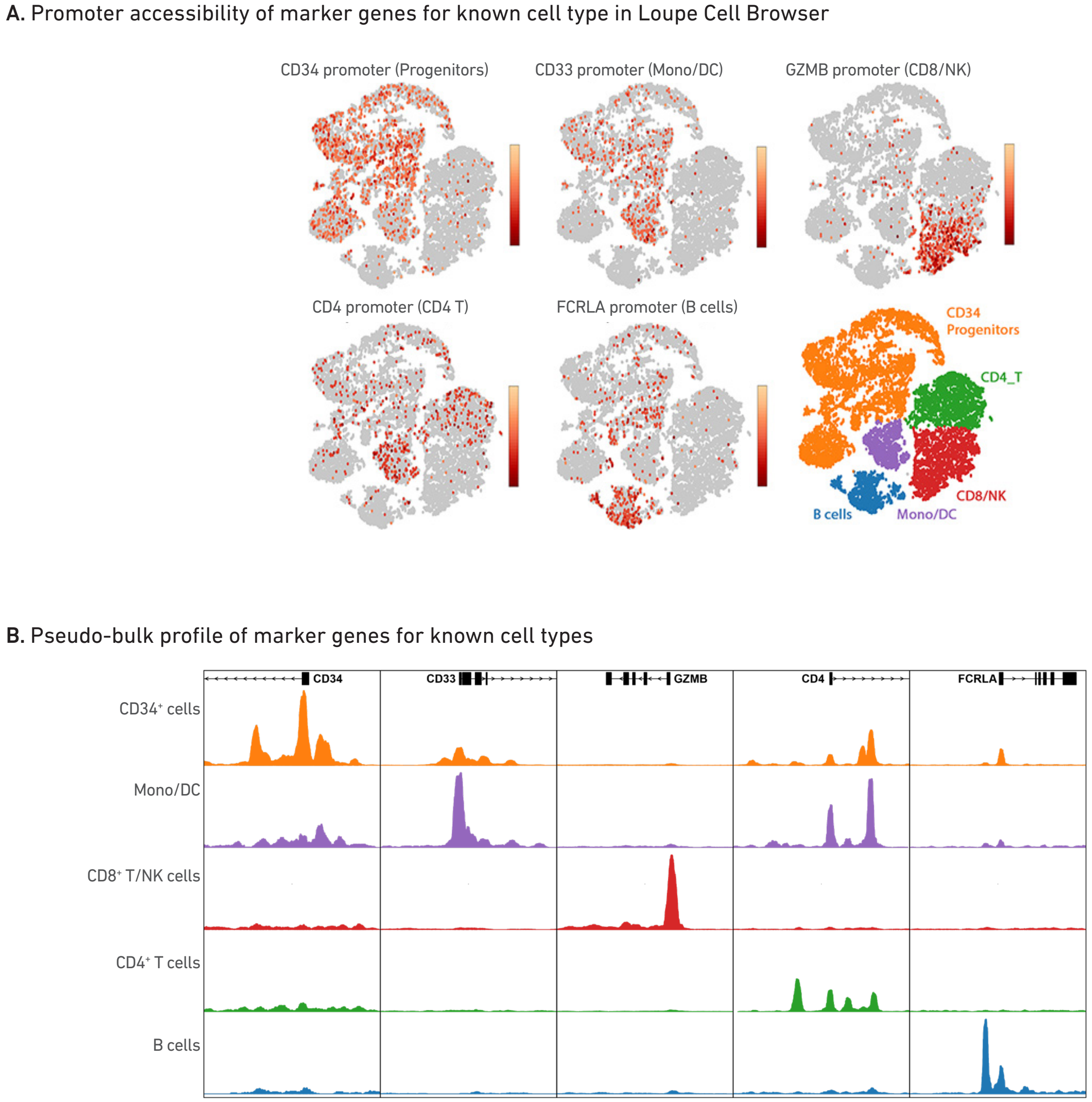
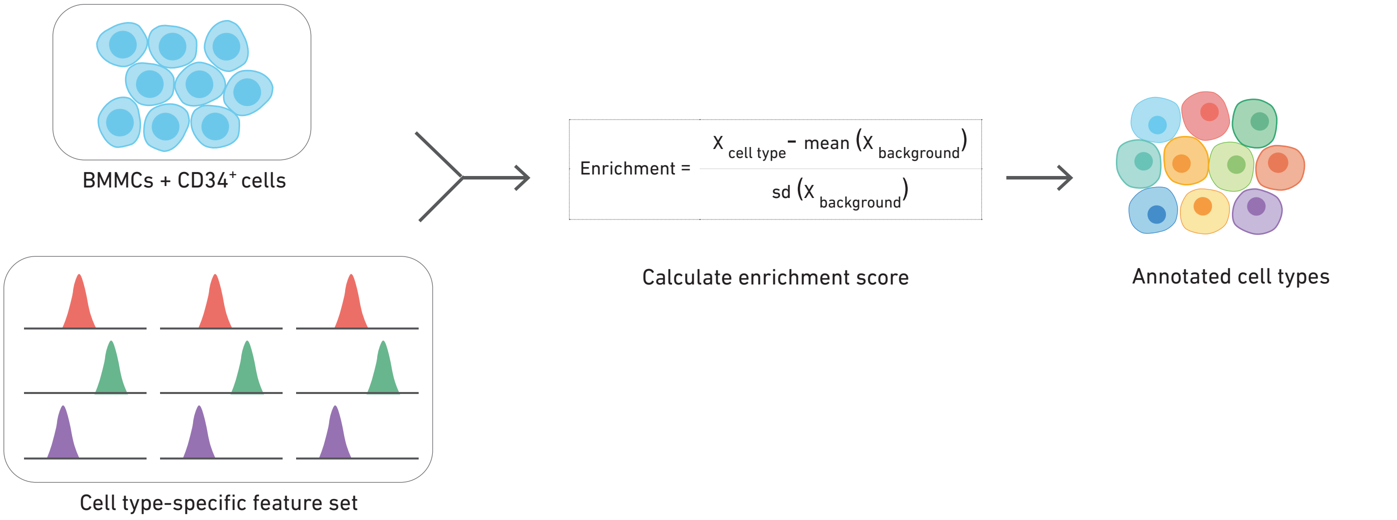


Figure 1. Annotating cell type using promoters. A. Promoter accessibility of marker genes for known cell types and subsequent cell type annotation. Colors indicate log transformed count of selected promoters, red = high values. **B.** Pseudo-bulk cut site distribution at promoters of the marker genes examined in **A.** Cut site bedgraph files were exported from Loupe Cell Browser. NK: Natural killer cells; Mono: Monocytes; DC: Dendritic cells

Strategy 2: Annotation Using Cell Type-Specific Feature Set

Methods



$$\text{Enrichment} = \frac{X_{\text{cell type}} - \text{mean}\left(X_{\text{background}}\right)}{\text{sd}\left(X_{\text{background}}\right)}$$

Results

A. Cell type enrichment score for selected cell types

B. Application of enrichment scoring scheme to BMMC + CD34^{+} single cell ATAC-seq data

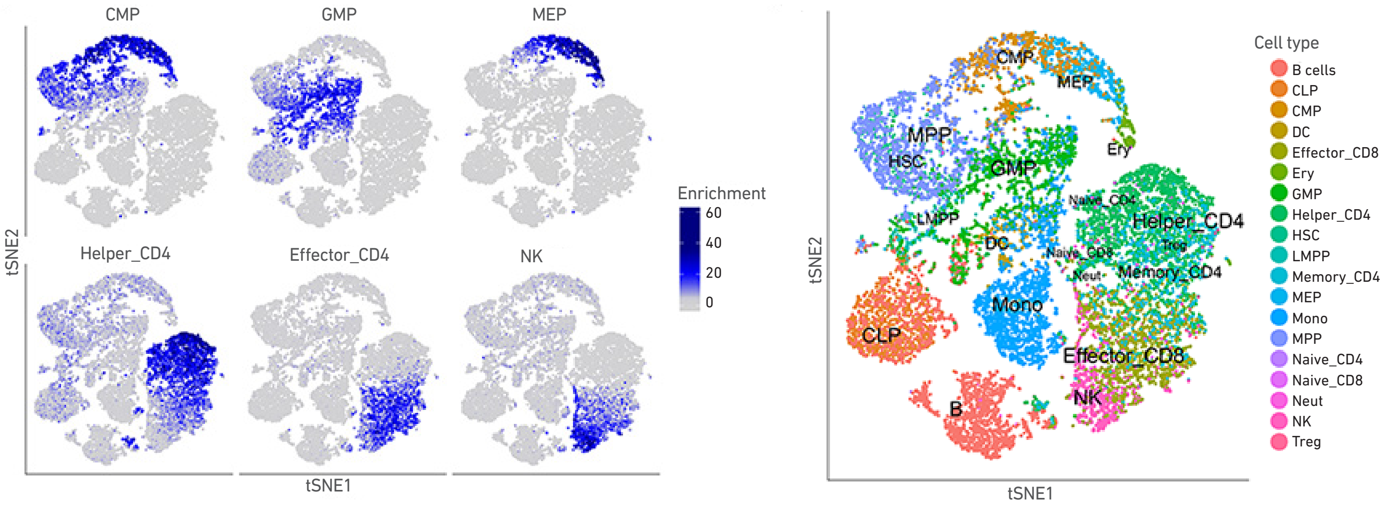


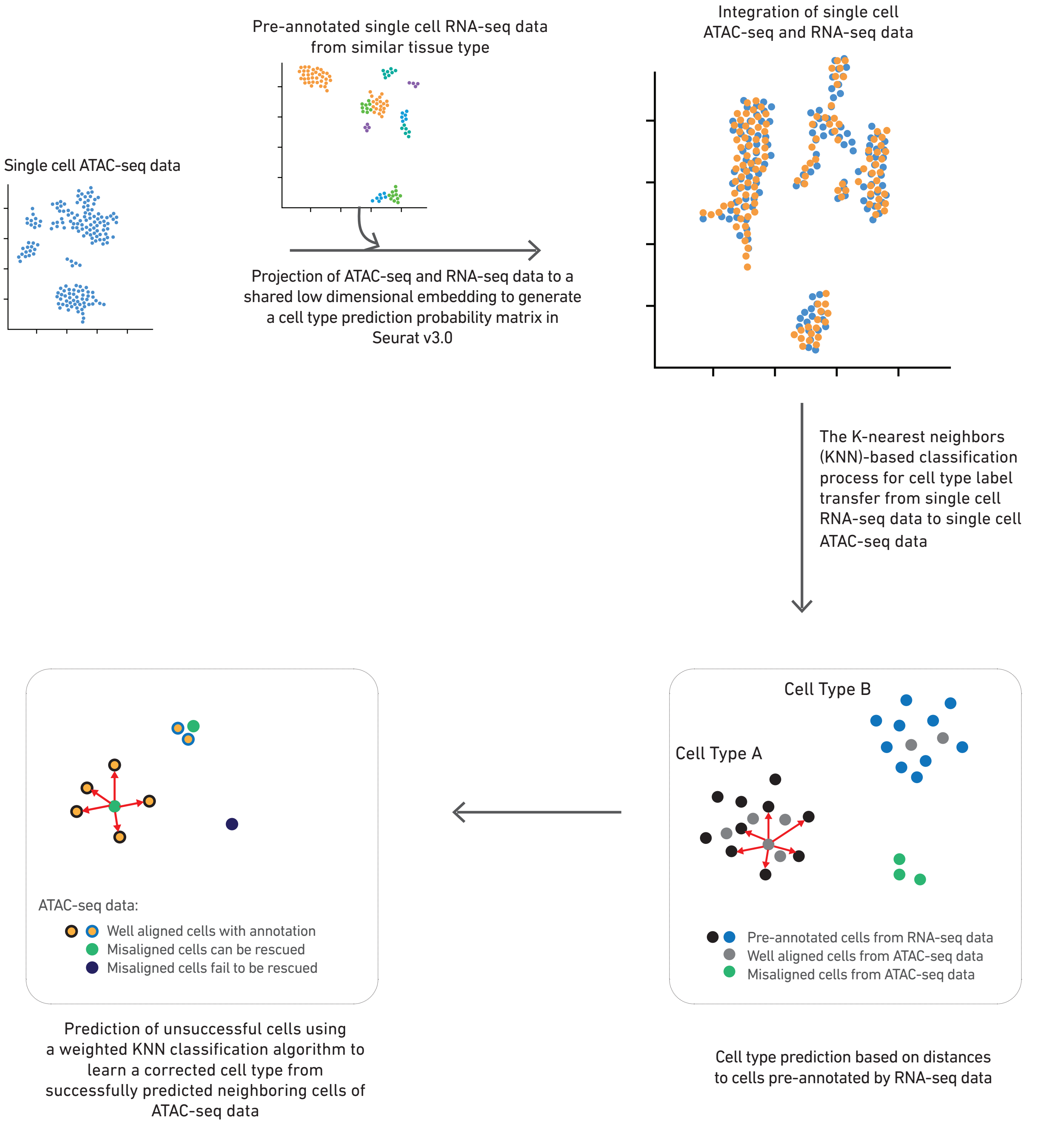
Figure 2. Annotating cell type using cell type-specific feature set. A. Distribution of cell type enrichment score for selected cell types. **B**. 19 major cell types were identified in BMMCs + CD34^{+} cells in the single cell ATAC-seq data. tSNE projections were obtained directly from Cell Ranger ATAC pipeline. Sizes of cell type labels are displayed prorated to the abundance of each type.

CLP: Common lymphoid progenitors
CMP: Common myeloid progenitor
DC: Dendritic cells
Ery: Erythroid
GMP: Granulocyte-macrophage progenitor
HSC: Hematopoietic stem cells

LMPP: Lympho-myeloid primed progenitor
MEP: Megakaryocyte-erythroid progenitor
Mono: Monocytes
MPPs: Multipotent progenitor cells
Neut: Neutrophills
NK: Natural killer cells

Strategy 3: Annotation Using RNA Sequencing Data as Reference

Methods


Cell Type B
Cell Type A
ATAC-seq data:
● ● Well aligned cells with annotation
● Misaligned cells can be rescued
● Misaligned cells fail to be rescued
● ● Pre-annotated cells from RNA-seq data
● Well aligned cells from ATAC-seq data
● Misaligned cells from ATAC-seq data

Single cell ATAC-seq data

Pre-annotated single cell RNA-seq data from similar tissue type

Projection of ATAC-seq and RNA-seq data to a shared low dimensional embedding to generate a cell type prediction probability matrix in Seurat v3.0

Integration of single cell ATAC-seq and RNA-seq data

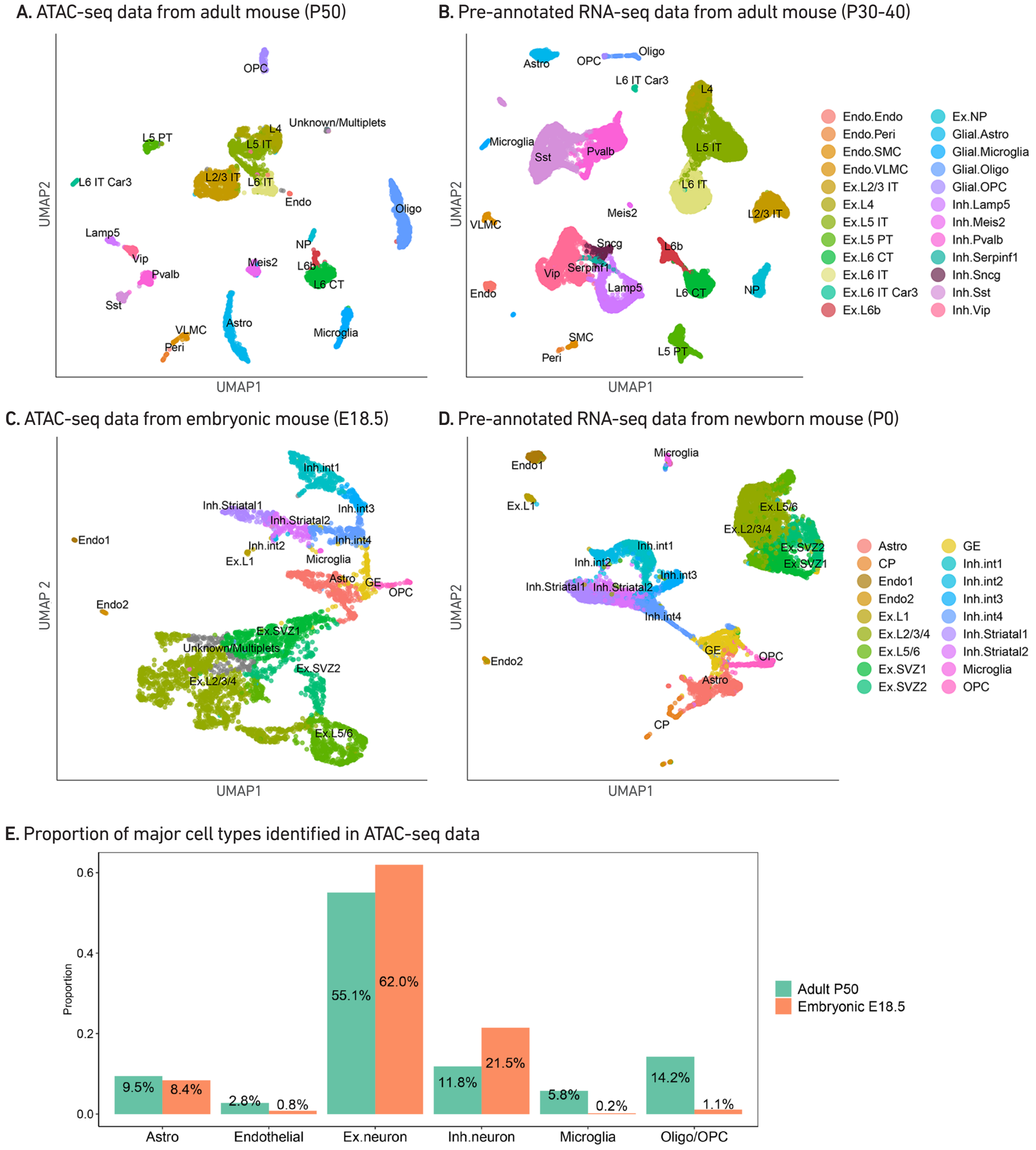
The K-nearest neighbors (KNN)-based classification process for cell type label transfer from single cell RNA-seq data to single cell ATAC-seq data

Cell type prediction based on distances to cells pre-annotated by RNA-seq data

Prediction of unsuccessful cells using a weighted KNN classification algorithm to learn a corrected cell type from successfully predicted neighboring cells of ATAC-seq data

Strategy 3: Annotation Using RNA Sequencing Data as Reference

Results



A. ATAC-seq data from adult mouse (P50)

B. Pre-annotated RNA-seq data from adult mouse (P30-40)

C. ATAC-seq data from embryonic mouse (E18.5)

D. Pre-annotated RNA-seq data from newborn mouse (P0)

E. Proportion of major cell types identified in ATAC-seq data

Figure 3. Annotation using RNA-seq data as reference. UMAP plots of ATAC-seq data from adult and embryonic mouse cortex annotated using pre-annotated RNA-seq data are shown in **A** and **C**, respectively. UMAP plots of the pre-annotated RNA-seq data from adult and newborn mouse cortex, are shown in **B** and **D**, respectively. The integration shows considerable overlap between the reference RNA-seq and the ATAC-seq data. More than 20 distinct cell types in adult mouse cortex and 18 major cell types in E18.5 mouse cortex tissue were identified.
E. Proportion of major cell types identified in adult and embryonic mouse brain cortex.

Astro: Astrocytes
CP: Choroid plexus
Endo: Endothelial cells
Ex: Excitatory neurons
GE: Ganglionic eminence
Inh: Inhibitory neurons
Int: Interneurons
Oligo: Oligodendrocytes
OPC: Oligodendrocyte progenitor cells
Peri: Pericytes
SMC: Smooth muscle cells
SVZ: Subventricular zone
VLMC: Vascular and leptomeningeal cells

Validation of Cell Type Annotation Using RNA Sequencing Data as Reference

Validation Using Gene Activity Scores

To validate cell type annotation, the R package Cicero (5) was applied to calculate the gene activity (GA) score of single cell ATAC-seq data from both embryonic and adult mouse tissue. For calculating the GA score, the peak-to-gene annotation and tSNE coordinates (as the reduced_coordinates) were obtained directly from the Cell Ranger ATAC output. Known markers of excitatory neurons, inhibitory neurons, and various glial cell types identified using strategy 3 (Figure 3C), were examined to confirm proper annotation (Figure 4A-B).

GA score distribution in UMAP single cell projections confirmed results of strategy 3. For example, accessibility of neuronal progenitor marker Eomes was higher in embryonic E18.5 brain compared to adult P50, with strong enrichment in the SVZ region (Figure 4A), thus validating the results of strategy 3.

Validation Using Transcription Factor (TF) Deviation Scores

Transcription factor (TF) deviation scores computed by chromVAR (6) measure TF activity and can be another source of validation of cell type annotations. To measure global TF activity, the input count matrix from the TF-barcode matrix of the Cell Ranger ATAC pipeline was obtained and the JASPAR motif database was selected as the input motif database. TF deviation scores of cell types identified in single cell ATAC-seq data from adult tissue using strategy 3 (Figure 3A) were then calculated using the recommended chromVAR workflow (Figure 4C).

Cell type-specific transcription factors, such as Noto in astrocytes and Spi1 in microglia showed exclusive activity in the corresponding cell types (Figure 4C). Comparison of Mef2c TF deviation scores in inhibitory neuron subtypes confirmed the previous reports of elevated activity of Mef2c in Pvalb subtype (7).

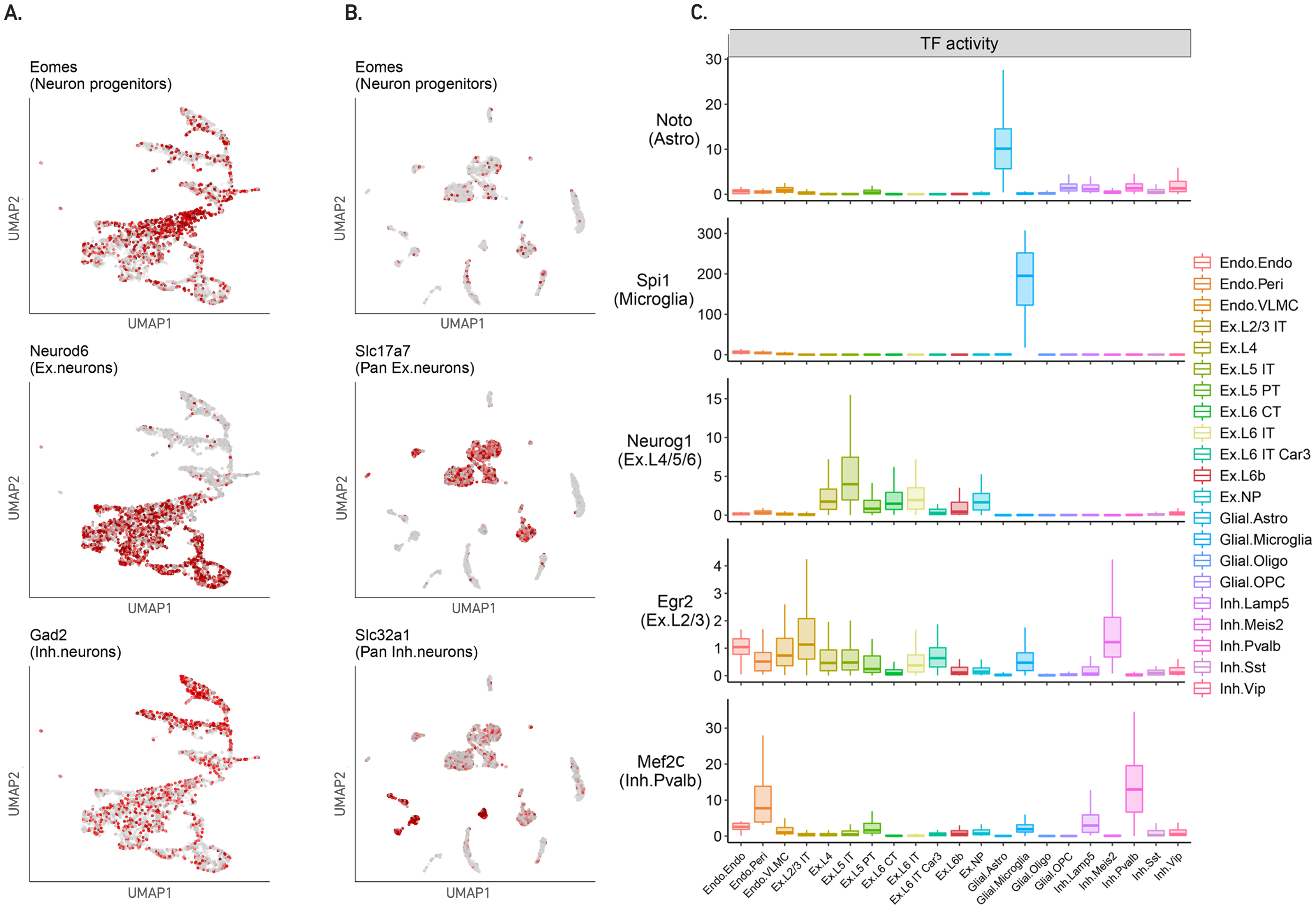


Figure 4. Validation of cell type annotation. A. E18.5 gene activity distribution in UMAP single cell projection. **B.** P50 adult gene activity distribution in UMAP single cell projection. Shade of red = high gene activity levels, gray = undetectable accessibility in promoter and nearby enhancers. **C.** Adult TF activity by cell types. Y axes are scaled transcription factor activity scores, based on the -log10 of p values converted from chromVAR TF deviation z scores.

Discussion

Cell type annotation using cell type-specific cis-regulatory elements showed a clear enrichment of promoter accessibility in different sub-populations of cells, allowing identification of the major cell types in bone marrow mononuclear cells. In the pseudo-bulk profile where all the cells in a cluster are aggregated into a single track, the chromatin accessibility near the promoter of gene markers showed a more complex pattern. For example, the CD4 promoter showed multiple enrichment peaks, only one of which had exclusive CD4$^+$ T-cell specificity, while other peaks also displayed strong accessibility in monocytes and stem cell populations.

The annotation of cell types using a cell type-specific feature set is an extension of the traditional gene marker-based strategy in which a list of marker genes is replaced by an interpretable feature set, thus providing flexibility to incorporate bulk data, transcription factor motif sites, or pre-annotated gene sets. The refined cell type annotation illustrated more details of the substructure of the CD34$^+$ progenitor population, including the multipotent stem cell population (HSC, MPP) and committed lineage progenitor cells (CMP. MEP, GMP and CLP) (Figure 2A-B). The substructure of the progenitor population can also be matched with terminally differentiated cells from different lineages to form a complete developmental trajectory, which is explored in more detail in the Application Note – Deciphering Epigenetic Regulation with Single Cell ATAC-Seq (LIT000055).

The unsupervised, integration-based strategy co-embeds single cell ATAC-seq data within reference single cell RNA-seq data and does not require any prior knowledge of marker genes. The annotation can be validated by computing gene and transcription factor activity scores (Figure 4A-C). The integration-based strategy can also be extended for annotation of any type of single cell data. For example, it can be easily adopted to the annotation of single cell RNA-seq data using pre-annotated single cell RNA datasets. Moreover, the integrated data provide a starting point for delineation of regulatory relationships between enhancer and target genes and ultimately gene regulatory networks.

Conclusion

In summary, three complementary cell type annotation strategies were demonstrated for single cell ATAC-seq data. The cell type annotation method chosen will depend on the knowledge or data available for the sample type of interest, or for a similar sample type. The first strategy, which uses known cell type markers, is the simplest and can be easily visualized in Loupe Cell Browser. The second and third strategies require additional bioinformatic processing and complementary reference datasets (e.g. bulk ATAC-seq or single cell RNA-seq).

References

1. M R Corces et al, Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* 48, 1193–1203 (2016).
2. L Loo et al, Single-cell transcriptomic analysis of mouse neocortical development. *Nat. Commun.* 10, 134 (2019).
3. B Tasic et al, Shared and distinct transcriptomic cell types across neocortical areas. *Nature.* 563, 72–78 (2018) .
4. T Stuart et al, Comprehensive integration of single-cell data. *Cell.* 177, 1888-1902 (2019).
5. H A Pliner et al, Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell.* 71, 858–871(2018).
6. A N Schep, B Wu, J D Buenrostro, W J Greenleaf, chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods.* 14, 975–978 (2017).
7. C Mayer et al, Developmental diversification of cortical inhibitory interneurons. *Nature.* 555, 457-462 (2018).

The use of 10x Genomics products in practicing the methods set forth herein has not been validated by 10x Genomics, and such non-validated use is NOT COVERED BY 10X GENOMICS STANDARD WARRANTY, AND 10X GENOMICS HEREBY DISCLAIMS ANY AND ALL WARRANTIES FOR SUCH USE. Nothing in this document should be construed as altering, waiving or amending in any manner 10x Genomics terms and conditions of sale for the Chromium Controller or the Chromium Single Cell Controller, consumables or software, including without limitation such terms and conditions relating to certain use restrictions, limited license, warranty and limitation of liability, and nothing in this document shall be deemed to be Documentation, as that term is set forth in such terms and conditions of sale. Nothing in this document shall be construed as any representation by 10x Genomics that it currently or will at any time in the future offer or in any way support any application set forth herein.

Contact:
support@10xgenomics.com
10x Genomics
6230 Stoneridge Mall Road
Pleasanton, CA 94588 USA