An Approach to Identify Expression Quantitative Super Enhancers Using Single-cell Multi-omic Data

Guoshuai Cai[1]

[1]*Department of Environmental Health Science, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29208, USA*

Corresponding to:

Guoshuai Cai, Ph.D.

Department of Environmental Health Science, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29208, USA

Discovery 449, 915 Greene St., Columbia, SC, 29208

Tel: 803-777-4120

Email: gcai@mailbox.sc.edu

Abstract

Super enhancers (SEs) drives cell identity and disease related genes. However, current methods for studying associations between SE and gene expression are time consuming, costly and have difficulties in scalability.

This study formulated an approach for screening genome-wide SE-expression associations using single-cell multi-omic data of transcriptome and H3K27ac histone modification. A pipeline was also constructed for an easy workflow application. Further application identified expression quantitative SEs (EqSEs) in brain and found they mark cell types. Moreover, our analysis provided new insights into the functional role of SEs close to *Zbtb20* and *Prkcb* in frontal cortex near-projecting neurons and hippocampus CA1 neurons, linking to neuron development and neurological diseases.

Collectively, this study provides a new tool for studying SE-expression associations and identifying significant expression associated SEs, which pave the way for understanding the regulatory role of SE in gene expression and related cellular and disease development.

Introduction

Super enhancers (SEs) are large clusters of enhancer elements, which have been found to be cell-type specific and activate the expression of cell identity genes. Compared to the typical enhancers, SEs bind to high levels of master transcription factors and Mediator coactivator and drives high-level expression of their associated genes [1]. Stimulating oncogenic transcription and other processes, SEs play key roles in the pathogenesis of various diseases, including cancers, neurodegeneration and others [2]. Therefore, understanding the regulatory role of SEs on gene expression will provide new insights into the cell and disease development. Genome-wide SE-expression associations will provide key data for this aim. However, current studies mainly rely on knock-out experiments on a single or a few SEs which are time consuming and costly, and have difficulties in scalability. In this study, a single-cell multi-omics based approach was developed to unblock this significant bottleneck.

The proposed approach studies SE-expression associations and identifies significant expression associated SEs (referred to as expression quantitative SEs, EqSEs) by analyzing single cell multi-omic data of transcriptome and H3K27ac histone modification (which is a reliable surrogate marker for SE identification [2]). In the following application study, EqSEs were identified and found to be activated in specific mouse brain cell types. The synchronized activity of SE and gene expression in frontal cortex (FC) and hippocampus (HC) neuron subtypes provide new evidence of the important role of SEs in neuron development and neurological diseases.

Methods

Datasets

Multi-omic data of H3K27ac modification and transcriptome of single cells from the frontal cortex and hippocampus tissues of Male C57BL/6J mice were generated by Paired-Tag [3]. The processed RNA sequencing read counts were downloaded from GEO GSE152020 and the raw protein binding DNA sequencing reads were downloaded from the SRA project of PRJNA638110.

Single-cell transcriptomic data processing and cell type identification

Single-cell transcriptomic data were processed and analyzed using Seurat [4]. The RNA sequencing read counts were normalized using the “LogNormalize” method and further centered and scaled by standard deviations. 2000 highly variable feature were selected, and dimension of their space was reduced using the top 20 principal components. Further, cell clusters were identified using the Louvain optimization-based clustering method and were visualized in reduced dimensions of Uniform Manifold Approximation and Projection (UMAP). Based on the expression of markers that were used in the original study [3], 18 cell types were identified, including 7 FC cell types (FC L2/3, FC L4, FC L5, FC PT, FC NP, FC CT, FC L6), 3 HC cell types (HC CA1, HC CA2/3, HC DG), 3 interneurons (InNeu-CGE, InNeu-Sst, InNeu-Pvalb), Oligodendrocyte

precursor cells (OPC), Oligodendrocytes (Oligo), Astrocytes, Endothelial cells and Ependymal cells. The cell clusters and corresponding marker expression profile were shown in Figure S1.

Single-cell histone modification data processing and SE identification

Single-cell protein binding DNA sequencing reads were processed following the method of Zhu et al. [3] for barcode extraction and genome mapping to the mm10 mouse genome assembly. Then, in cells of each major neuron type of FC, HC and InNeu which were identified by above transcriptome analysis, the peaks of H3K27ac signal were identified by MACS2 [5]. From these peaks, SEs were further identified by ROSE [1] in each single cell and H3K27ac binding DNA sequencing reads at SE regions were counted for quantification. In a particular cell, the signals were normalized by all reads with a consistent scale factor, 10000, and log transformation was applied to transform data to approximately conform to normality.

EqSE identification

For each neuron type, the associations between the H3K27ac signal at each SE and the expression of its proximal gene was assessed by Pearson correlation coefficients (r). To reduce the bias and noise of single-cell data due to its high drop-out rate, cells with missing values were filtered out from correlation calculation. To obtain reliable correlation assessment, only SE-gene pairs that have non-missing multi-modal data in more than 300 cells and more than 20 in each cell type of FC, HC and InNeu were analyzed in this study. EqSEs were considered to have a notable correlation with $r > 0.2$ in at least one cell type.

Results

A workflow/pipeline for screening expression quantitative SEs

To enable the EqSE screening, a workflow was designed as illustrated in Figure 1. It is a multi-step procedure for data processing, sequentially with demultiplexing, transcriptome/genome mapping, gene expression quantification, clustering and cell type identification based on single-cell transcriptome data, histone modification peak calling, SE identification, histone modification quantification, multi-modal data conjugation, and EqSE identification. The details and tool implementation were described in Methods. The outputs of this workflow are cell type-specific super enhancer regions, single-cell gene expression and histone modification levels, SE-expression association, and identified EqSEs. This new approach leverages single-cell multiomic data of three key advantages: (1) the same-cell joint profiling of transcriptome and histone modification captures cell-to-cell variation which was ignored by traditional bulk data, (2) the large scale of single-cell measurements provides favorable statistical power in association assessment, which was a critical challenge for traditional bulk tissue experiments, and (3) the single-cell resolution provides an opportunity to build SE atlas in atissue which has been known to be cell type-specific. In support of the workflow application, a pipeline was constructed and is available in Github https://github.com/thecailab/EqSE.

EqSE activeness are cell type-specific in brain

Following the workflow, this study assessed the SE-expression association in three major neuron types, FC, HC and interneurons (InNeu). Consistent associations (i.e., correlations) were widely found cross these neuron types (Fig. 2). 25 EqSEs were identified ($r > 0.2$, see Methods) and the expression of 16 associated genes statistically significantly mark brain cell types. As shown in Figure 3A, specific expression was found on *Kcnip4* and *A830082K12Rik* in FC L2/3, L4 and L5; *Osbpl1a*, *Phactr1*, *Prickle2* and *Prkcb* in FC and HC; *Sh3rf3*, *Klhl29*, *Hivep2* and *Pde4dip* in neuron cells; *Zbtb20* in HC and non-neuron cells; *Zbtb16* in HC and InNeu-CGE; *Tiam1* and *Scn3a* in HC DG; *Lrrc4* in InNeu-Pvalb; *Ctnna3* in Oligo and Ependymal cells. Despites of low

115 coverage, histone modification at these EqSEs also showed a corresponding pattern of
116 activeness (Fig. 3B).

117 Interestingly, the histone modification at SEs close to *Zbtb20* and *Prkcb* are specifically
118 correlated to gene expression in FC and HC respectively (Figs. 2 and 4). As compared to other
119 subtypes of FC neurons, synchronized activeness of histone modification at SE and gene
120 expression of *Zbtb20* were found in FC near-projecting neurons (FC NP) with relatively higher
121 detection rate and average level (Fig.4 left, p-value= 9.86×10^{-6} for gene expression, non-
122 significant for histone modification data due to the low sequencing coverage). In HC CA1 neurons,
123 significant activeness of *Prkcb* were observed on both histone modification at SE (p-value=
124 9.66×10^{-3}) and gene expression (p-value= 8.52×10^{-58}), as compared to other HC neurons (Fig.4
125 right).

126

127 **Discussion**

128 This study developed a novel workflow with pipeline to study SE-expression association and
129 identify EqSEs. It provides a new approach for interactively analyzing the same-cell transcriptome
130 and histone modification and investigating SE-related regulatory programs in cell populations. Its
131 performance was cross validated by the broad consistence of associations in three major neuron
132 cell types. The study also identified specific associations at *Zbtb20* in FC and at *Prkcb* in HC,
133 indicating SE-related regulation could be cell type-specific. Previous studies suggested an
134 important role of *Zbtb20* in cell fate specification and in instructing the generation of a subset of
135 callosal projections neurons in cortical layers II/III [6, 7]. Aligning well with these studies, active
136 gene expression of *Zbtb20* was found in FC L2/3 and NP cells. In addition, active expression of
137 *Prkcb* was found in HC CA1 cells, in consistent with the previous findings that CA1 pyramidal
138 neurons is highly vulnerable in Alzheimer's disease (AD) and was characterized by *Prkcb*, a

139 causative factor of AD [8, 9]. In this study, the newly identified synchronous activation of SEs in
140 specific cell types may play critical roles in underlying mechanisms of these related diseases.

141 A limitation resides in the low per cell sequencing coverage compromises the sensitivity
142 of the single-cell profiles, especially for histone modification detection. This limitation will be
143 remediated by the reducing cost of sequencing and single cell library construction. With necessary
144 modifications, the proposed workflow is expected to be easily extended for data from future
145 improved platforms, and other platforms for assessing the associations of other features, such as
146 methylation and gene expression.

147

148 *Conflict of Interest: none declared.*

149

150 **References**

151 1. Whyte, W.A., et al., *Master transcription factors and mediator establish super-enhancers at key*
152 *cell identity genes.* Cell, 2013. **153**(2): p. 307-19.
153 2. Hnisz, D., et al., *Super-enhancers in the control of cell identity and disease.* Cell, 2013. **155**(4): p.
154 934-47.
155 3. Zhu, C., et al., *Joint profiling of histone modifications and transcriptome in single cells from*
156 *mouse brain.* Nat Methods, 2021. **18**(3): p. 283-292.
157 4. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902
158 e21.
159 5. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.
160 6. Medeiros de Araujo, J.A., et al., *ZBTB20 is crucial for the specification of a subset of callosal*
161 *projection neurons and astrocytes in the mammalian neocortex.* Development, 2021. **148**(16).
162 7. Tonchev, A.B., et al., *Zbtb20 modulates the sequential generation of neuronal layers in*
163 *developing cortex.* Mol Brain, 2016. **9**(1): p. 65.
164 8. Gerschutz, A., et al., *Neuron-specific alterations in signal transduction pathways associated with*
165 *Alzheimer's disease.* J Alzheimers Dis, 2014. **40**(1): p. 135-42.
166 9. Zhou, Z., et al., *Molecular identification of protein kinase C beta in Alzheimer's disease.* Aging
167 (Albany NY), 2020. **12**(21): p. 21798-21808.

168

169

Figures

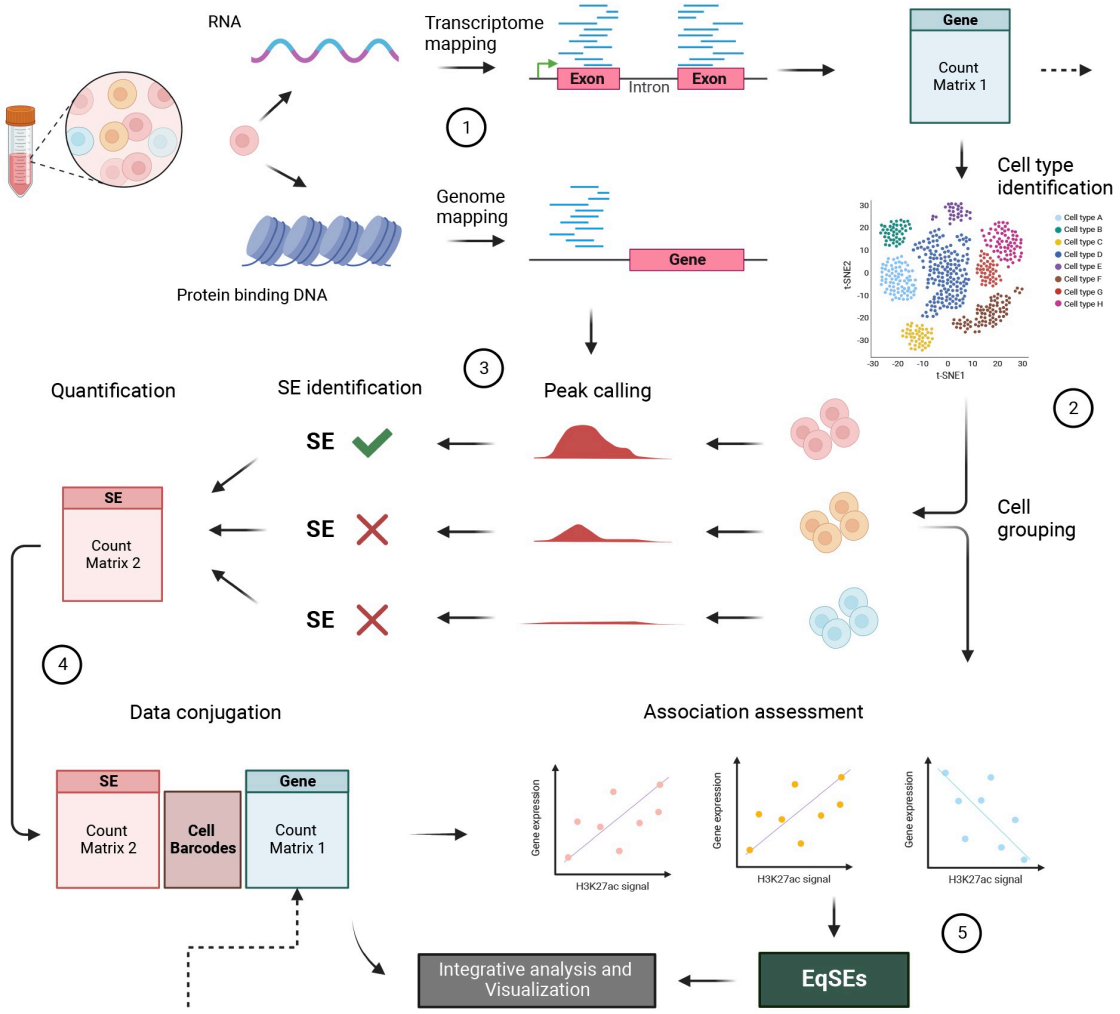

Figure 1. The workflow of EqSE identification using single-cell multiomic data. It constructs a multi-step processing of (1) sequencing data processing, including demultiplexing, transcriptome/genome mapping and gene expression quantification, (2) cell type identification based on single-cell transcriptomics, (3) SE identification and quantification based on cell type-level histone modification data, (4) multi-modal data conjugation, and (5) EqSE identification.

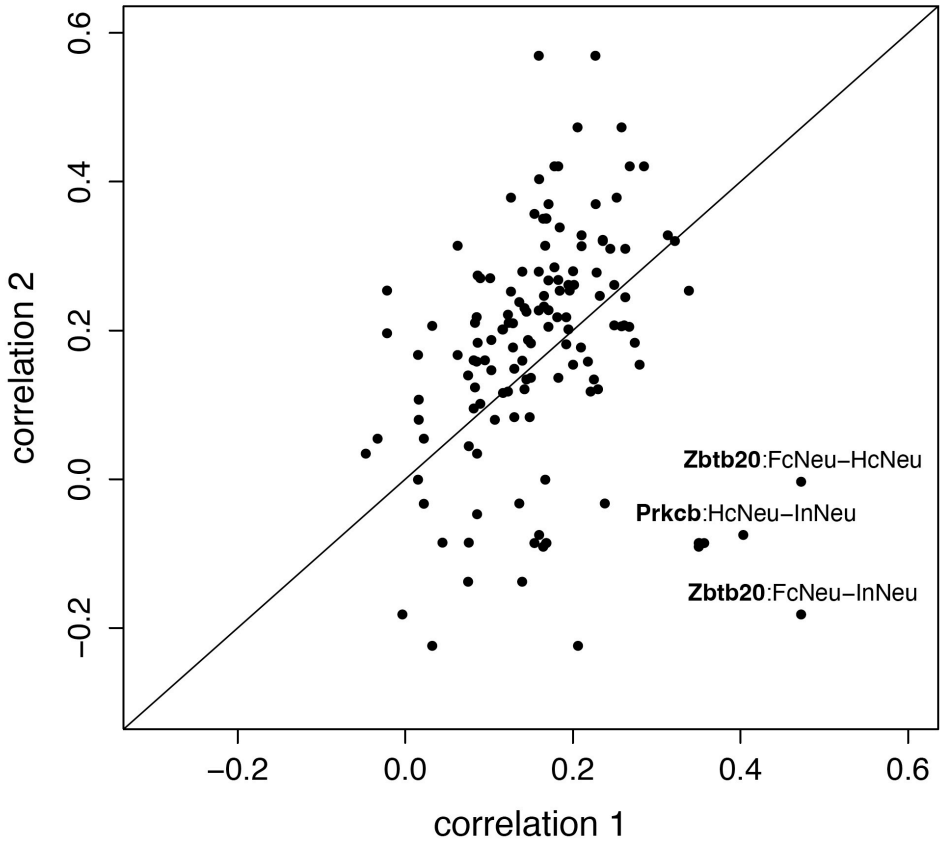

Figure 2. Comparison of SE-expression associations in Neuron subtypes. For each gene, correlation 1 and 2 are from FC, HC or InNeu for pairwise comparisons. The significant difference of SE-expression associations in *Zbtb20* between FC and HC/InNeu and *Prkcb* between HC and InNeu were identified and labeled above corresponding dots.

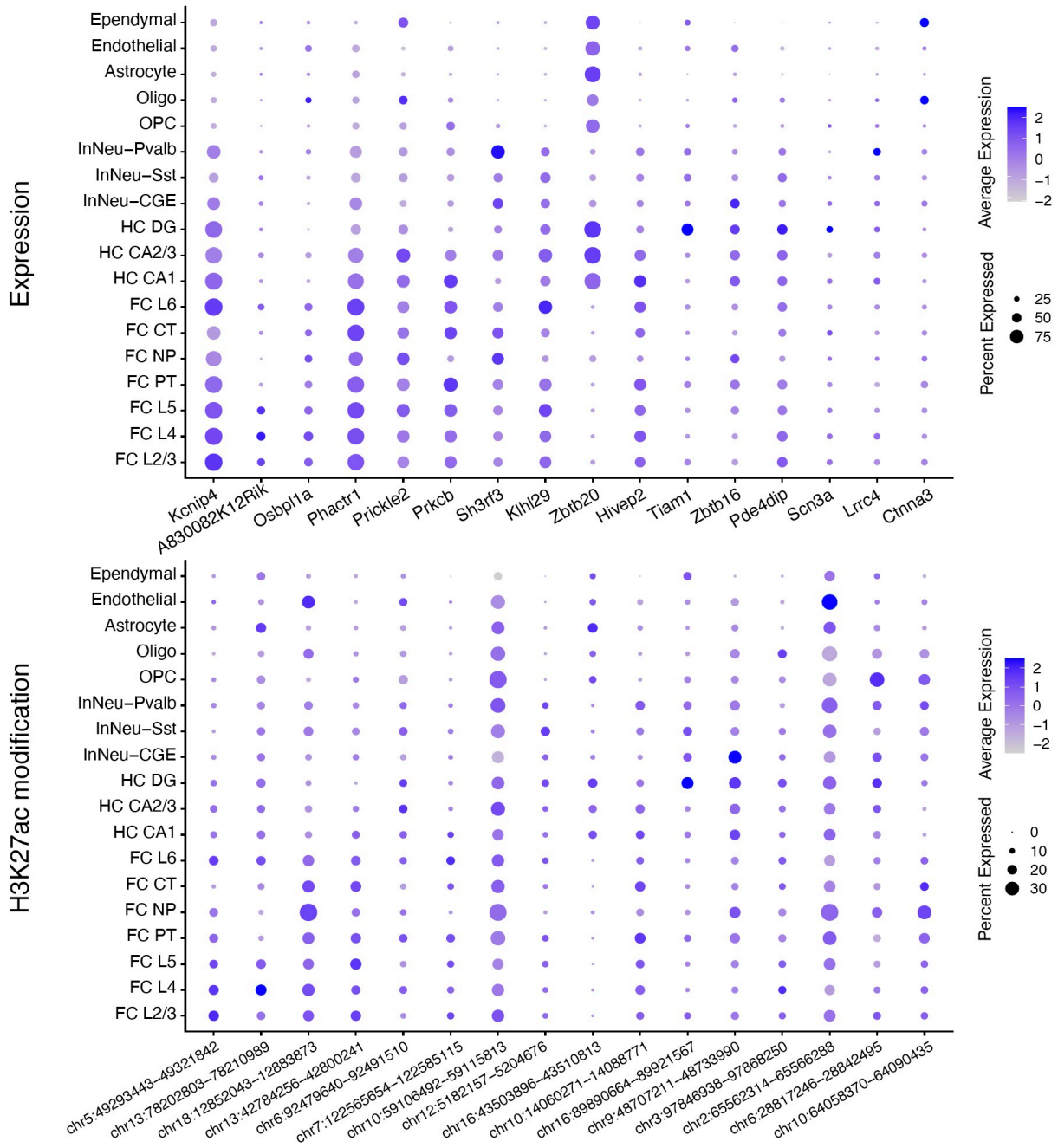

Figure 3. Profiles of EqSE features in brain cell types. In a particular cell type, the level of expression of genes (upper) and histone modification at each mapped EqSE (down) were shown by dot size and color. A larger dot size indicates a higher detection rate of gene expression within each cell type, while a darker color indicates a higher average level of signal.

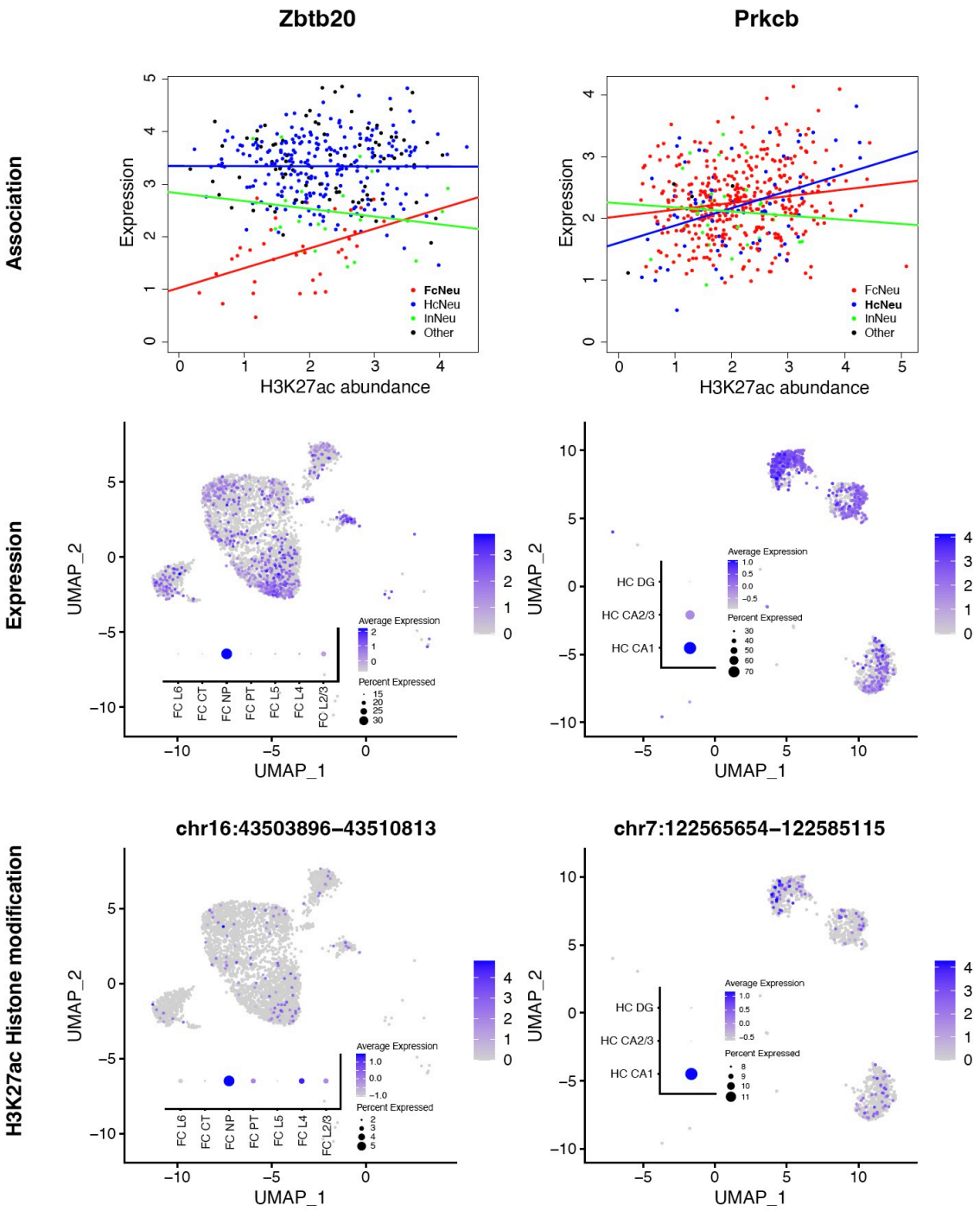

Figure 4. *Zbtb20* and *Prkcb* expression and histone modification at proximal SEs in neuron subtypes. The top shows the SE-expression correlation of *Zbtb20* and *Prkcb* in different neuron types. Uniform manifold approximation and projection (UMAP) plots of single-cell transcriptome with gene expression and histone modification are shown in middle and bottom respectively, for Zbtb20 in FC subtypes (left) and Prkcb in HC subtypes (right). Overlaid dot plots show average detection rates (corresponding to dot size, larger dot indicates higher detection rate) and average signal levels (corresponding to dot color, darker indicates higher signal level).

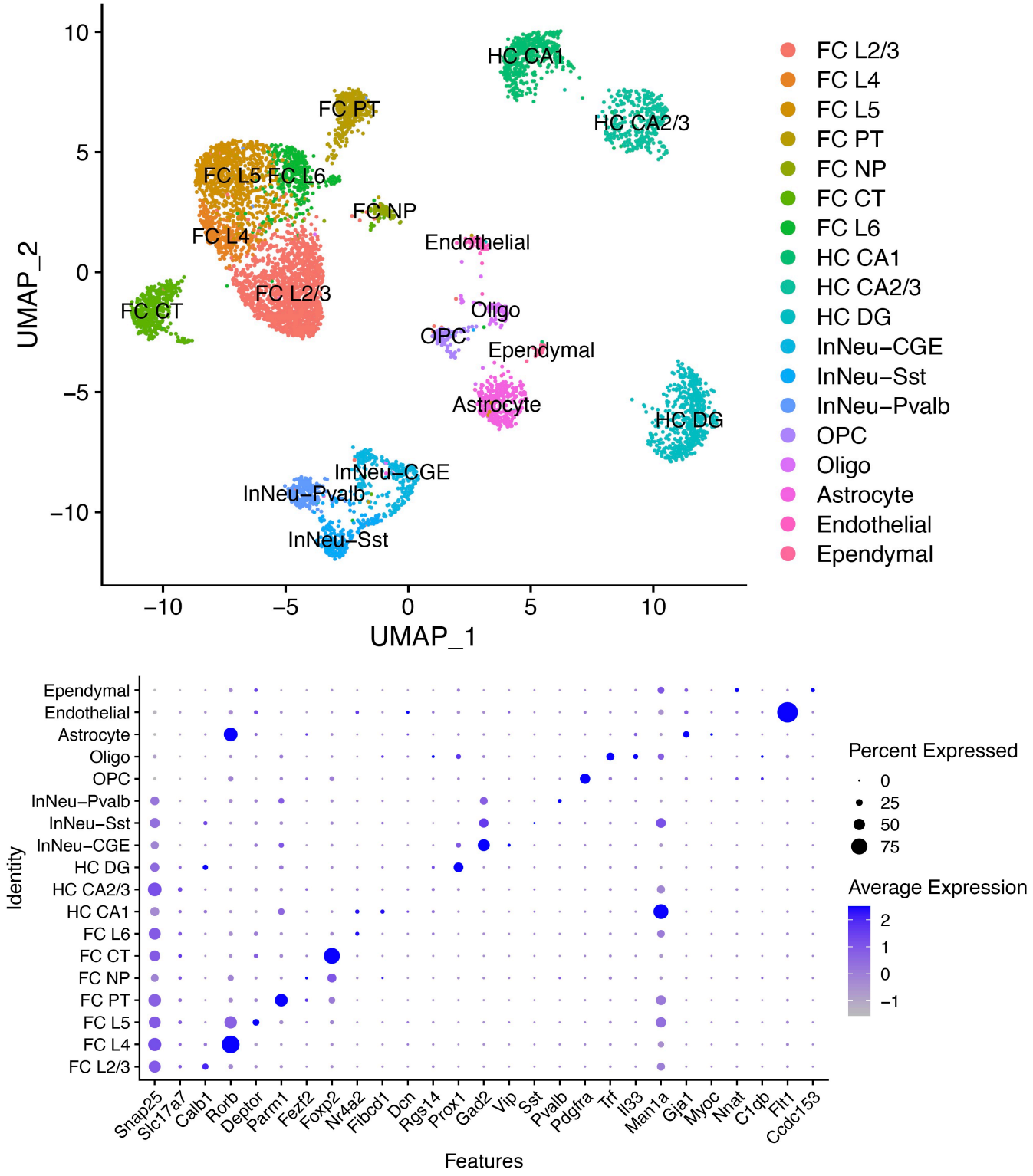

Figure S1. Brain cell clusters and marker expression profiles. For each cell type shown in the UMAP plot (top), the detection rate (corresponding to dot size, larger dot indicates higher detection rate) and average expression level (corresponding to dot color, darker indicates higher expression level) of each cell-type marker are shown in the dot plot (bottom).