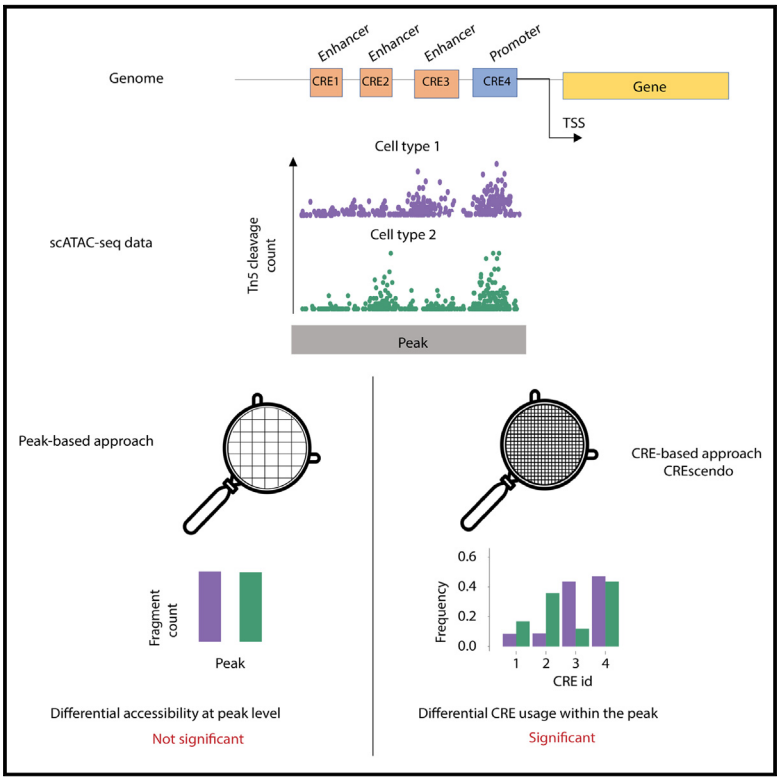
Cell Genomics

Technology

Capturing cell-type-specific activities of *cis*-regulatory elements from peak-based single-cell ATAC-seq

Graphical abstract



Highlights

- Arbitrary peaks mask cell-type-specific regulatory signals
- CREscendo uses Tn5 cleavage and CRE data to refine peak decomposition
- CRE-centric quantification improves precision, interpretability, and reproducibility

Authors

Mengjie Chen

Correspondence

mengjiechen@uchicago.edu

In brief

Mengjie Chen introduces CREscendo, a framework using Tn5 cleavage and ENCODE CREs to refine scATAC-seq peaks. Applied to PBMCs and mouse cortex, it recovers key regulatory signals that are missed by standard methods. This study advocates shifting from peak-based to CRE-centric analysis, enhancing precision and reproducibility.

Chen, 2025, Cell Genomics *5*, 100806
March 12, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.xgen.2025.100806

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

Capturing cell-type-specific activities of *cis*-regulatory elements from peak-based single-cell ATAC-seq

Mengjie Chen[1,2,*]
[1]Department of Medicine, Department of Human Genetics, and Department of Statistics, University of Chicago, Chicago, IL 60637, USA
[2]Lead contact
*Correspondence: mengjiechen@uchicago.edu
https://doi.org/10.1016/j.xgen.2025.100806

SUMMARY

Single-cell ATAC sequencing (scATAC-seq), a state-of-the-art genomic technique designed to map chromatin accessibility at the single-cell level, presents unique analytical challenges due to limited sampling and data sparsity. In this study, we use case studies to highlight the limitations of conventional peak-based methods for processing scATAC-seq data. These methods can fail to capture precise cell-type-specific regulatory signals, producing results that are difficult to interpret and lack portability, thereby compromising the reproducibility of research findings. To overcome these issues, we introduce CREscendo, a method that utilizes Tn5 cleavage frequencies and regulatory annotations to identify differential usage of candidate regulatory elements (CREs) across cell types. Our research advocates for moving away from traditional peak-based quantification in scATAC-seq toward a more robust framework that relies on a standardized reference of annotated CREs, enhancing both the accuracy and reproducibility of genomic studies.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) are pivotal for profiling gene-expression-level differences between cells and pinpointing the regulatory mechanisms that control their activity. These methodologies differ significantly in the extent of the genomic regions they target. scRNA-seq, especially in leading platforms like 10× Genomics, primarily captures transcribed RNAs, focusing on 100–300 nt upstream of the 3′ untranslated region (UTR) of transcripts. This results in a target size of approximately 6 million base pairs, assuming predominant 3′ UTR expression across (∼30,000) protein-coding genes. Conversely, scATAC-seq targets the entire expanse of open chromatin using a transposase enzyme called Tn5 to fragment and tag DNA from accessible chromatin regions.[1] Open chromatin comprises about 1%–2% of the genome in a given cell, translating to a target size of 30–60 million base pairs. Both approaches face significant analytical challenges due to limited sampling and data sparsity, but these issues are particularly pronounced in scATAC-seq.

Unlike RNA-seq reads, which map to a specific region of the gene, chromatin accessibility "peaks" are defined arbitrarily, making it difficult to consistently interpret scATAC-seq data or compare results from different scATAC-seq studies. A chromatin accessibility peak is essentially a genomic interval that shows an enrichment of sequencing reads compared to background levels.[2] The relatively high abundance of reads is interpreted as indicating greater Tn5 activity, potentially due to active candidate regulatory elements (CREs). Peak position and width are heavily influenced by methodological choices, yet the read counts within these intervals serve as the primary units for downstream analyses,[3–5] including clustering, differential accessibility (DA), and transcriptional factor motif enrichment analyses.[6–15]

In this manuscript, we present case studies that illustrate the challenges posed by standard approaches for processing scATAC-seq data, particularly for downstream analyses and interpretation. Using paired scRNA-seq and scATAC-seq (10× Multiome) data and scATAC-seq data from two independently collected 10× peripheral blood mononuclear cell (PBMC) datasets, we demonstrate significant variability in the chromatin accessibility peaks identified by three popular peak calling programs: Cell Ranger,[16] MACS2,[17] and MACS3. Strikingly, the majority of peaks identified by these approaches contain multiple CREs, and many fail to capture active CREs. In response to these concerns, we introduce CREscendo, a novel framework that identifies CRE-level signals that conventional peak-based scATAC-seq analyses often overlook. We also corroborate some of our findings using ENCODE data and discuss viable alternatives to the standard peak-based approach.

DESIGN

We developed the CREscendo framework to address the need for more standardized analysis methods in single-cell genomics. Variability in scATAC-seq data generation and analysis approaches makes it challenging to draw consistent conclusions about CRE usage, particularly when comparing chromatin accessibility data from different cell types or experiments. To

This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

illustrate the technical and methodological limitations of peak-based approaches, we analyzed two independent ATAC datasets collected from human PBMCs and one dataset collected from mouse cortex. The first PBMC dataset was generated using the 10× Multiome platform, which captures paired scRNA-seq and scATAC-seq data from the same cells. Focusing our analysis on the two most abundant cell types, $CD14^+$ monocytes and $CD8^+$ naive T cells, we directly compared the proportion of transcriptome coverage obtained by scRNA-seq to the proportion of genome coverage obtained by scATAC-seq. We then assessed the impact of data sparsity on peak calling, finding that differential CRE usage between cell types may frequently be obscured by wide peaks. These results were validated in our separate analysis of 10× Chromium X2 scATAC-seq data from $CD14^+$ monocytes and $CD4^+$ memory T cells (selected based on their prevalence and overlap with the multiome dataset) and in our analysis of mouse cortical cells (selected to determine whether our approach is effective for tissues other than blood). As described in detail below, our findings reveal important limitations of peak-based methods and demonstrate that these issues are prevalent across different scATAC-seq platforms and peak calling methods, including those that have been "improved."

RESULTS

Standard peak calling methods are strongly affected by data sparsity and largely fail to pinpoint individual CREs

Given the target size estimates, achieving comparable coverage between scATAC-seq and scRNA-seq would require scATAC-seq to have a library size nearly ten times larger than that of scRNA-seq. However, typical scATAC-seq library sizes are only about twice that of scRNA-seq, which we observed to be the case for the 10× Multiome dataset. In this example, median library sizes are 10,000 fragments for scATAC-seq vs. 5,000 unique molecular identifiers (UMIs) for scRNA-seq in $CD14^+$ and $CD8^+$ cells (Figure 1A). Consequently, scATAC-seq provides sparser coverage for broader targeted regions, which results in a shallower average fragment count per cell. When examining scRNA-seq, median UMI counts relative to gene sparsity show a dynamic range from 0 to 10 (Figures 1B and S1–S3). In contrast, scATAC-seq displays higher peak sparsity with narrower dynamic ranges, where 98.7% of these peak regions average fewer than one fragment per cell. Standard practices often involve binarizing ATAC fragment counts to assess accessibility; however, this can introduce errors, as zeros frequently represent inadequate sampling rather than true inaccessibility. Notably, a strong correlation exists between fragment counts and peak widths, suggesting that peak width could confound quantitative analyses (Figures 1C and S4).

Considering that peaks are expected to capture key regulatory elements like promoters, enhancers, and insulators,[18] we analyzed peak sizes against annotated candidate CREs (cCREs). Specifically, we used the Registry of cCREs v.3 from SCREEN provided by ENCODE,[19] which includes 1,063,878 cCREs identified by segregating signals across all biospecimens sourced from 1,518 cell types. cCREs are further categorized into candidate promoters, proximal enhancers, and distal enhancers and are annotated by CTCF binding status. Within the 10× Multiome PBMC dataset, we observed that the peak calling software Cell Ranger identifies larger peaks with a median size of 659nt, while MACS2, another popular tool, reports smaller median peak sizes of 502 nt. Both types of peaks generally exceed the median size of individual cCREs, which are 288 nt for enhancers and 328 nt for promoters (Figure 1D). Over 55.2% of peaks identified by MACS2 cover at least three CREs, and this number rises to 75% at transcription start sites (TSSs).

Focusing on MACS2 peaks, which are smaller than those identified by Cell Ranger, we next examined the frequency of Tn5 cleavage sites around peaks containing multiple CREs. Strikingly, we found that although peaks containing multiple CREs span a total of 53,588,484 nt, only 25.1% coincide with Tn5 cleavage sites. Within these cleavage sites, a significant 78.8% align with annotated CREs (Figure 1E). This finding is consistent when evaluating the Tn5 cleavage site ratio within each peak, and their proportion aligns with CREs at the cell type level as well (Figure 1F).

A similar pattern emerged when analyzing the mouse cortex 10× Chromium X2 scATAC-seq dataset. In this dataset, peaks containing multiple CREs span a total of 31,811,858 nt, yet only 16.9% overlap with Tn5 cleavage sites. Of these overlapping regions, a significant 71.5% align with annotated CREs. Collectively, these findings highlight that the peaks identified in scATAC-seq, regardless of the calling method, tend to exceed the size of CREs. However, only a modest portion of peak widths directly overlap with Tn5 cleavage sites.

Our analysis of the PBMC Chromium X2 scATAC-seq dataset (Figure S5) also yielded similar results, even when using a more recent peak caller, MACS3. Although MACS3 peaks tended to be smaller than those identified by Cell Ranger (median of 457 nt compared to 878 nt for Cell Ranger), they still span a wide range, with a median peak width greatly exceeding the average CRE length.

Together, our analyses suggest that, regardless of the calling method used, chromatin accessibility peaks from scATAC-seq data largely fail to capture granular regulatory elements. Instead, these peaks frequently include the intervals of adjacent CREs, particularly within TSSs, where CREs are densely situated. Our results highlight the limitations of peak calling for identifying critical *cis*-regulatory elements and suggest that peak-based approaches could be improved by the use of regulatory annotations.

Integrating CRE annotations into the scATAC-seq analysis pipeline improves interpretation of cell-type-specific regulatory features

We reasoned that functional annotations, such as CRE annotations from ENCODE and gene activity scores, could be leveraged to improve the interpretation of chromatin accessibility peaks. Like peak-level summaries, gene activity scores are key metrics for interpreting scATAC-seq data.[2] In this approach, a gene's transcriptional activity is estimated using all fragments mapping to the gene and its promoter (2 kb upstream plus the gene body), with higher scores indicating greater accessibility and increased transcriptional potential. We used gene activity scores and CRE annotations from ENCODE to interpret

CellPress
OPEN ACCESS

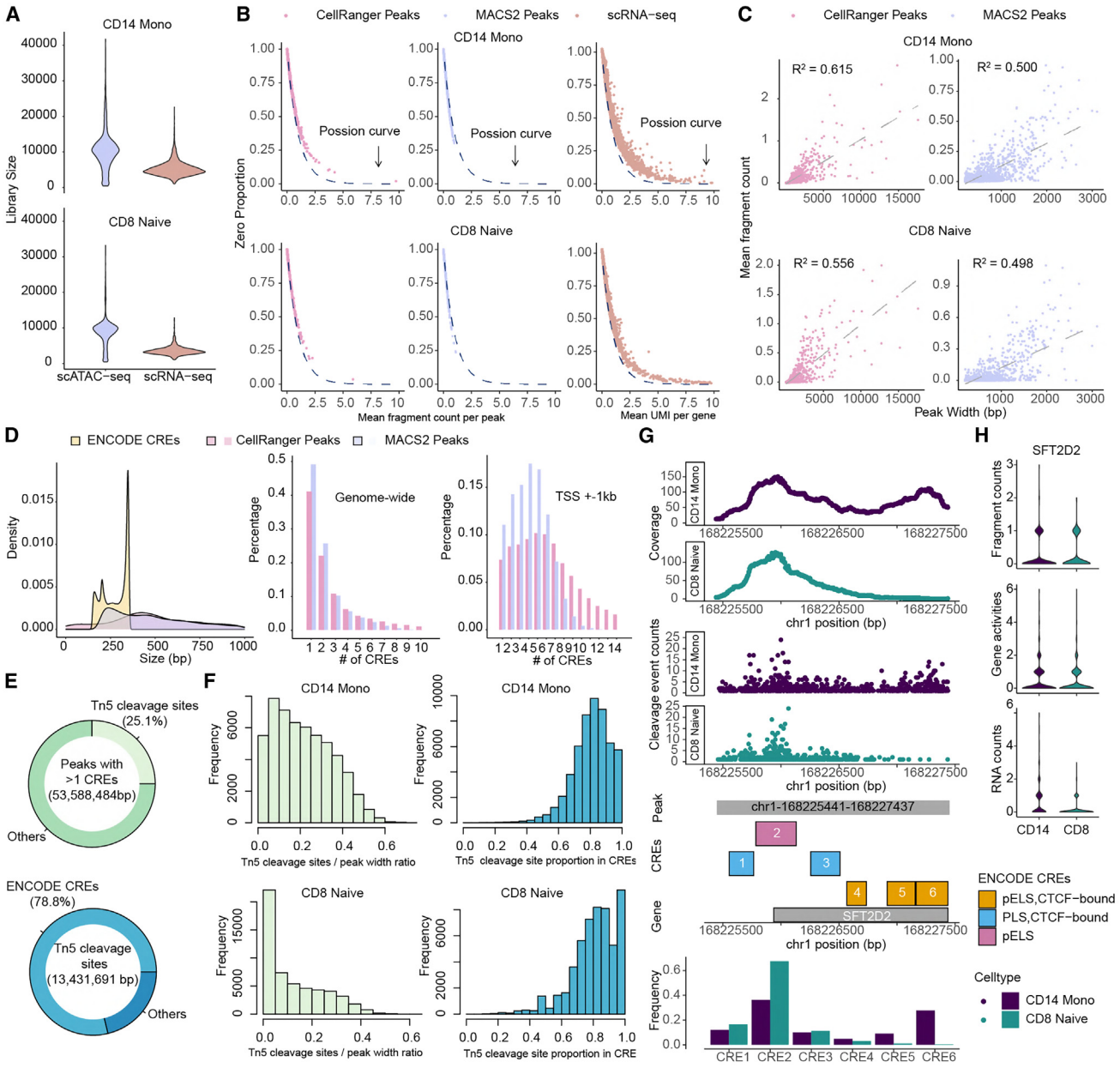


Figure 1. Characterization of peak-based analysis in PBMC 10× Multiome dataset for CD14$^+$ monocytes and CD8$^+$ naive T cells

(A) Comparison of library sizes between scATAC-seq and scRNA-seq.

(B) Proportions of zero counts plotted against mean fragment per peak (ATAC) and mean UMI per gene (RNA), with an assumed Poisson distribution curve shown as a dashed line.

(C) Relationship between mean fragment count and peak width.

(D) Density plots comparing sizes of peaks and CREs annotated by ENCODE, including the percentage of peaks covering multiple CREs genome wide and within ±1 kb of TSSs.

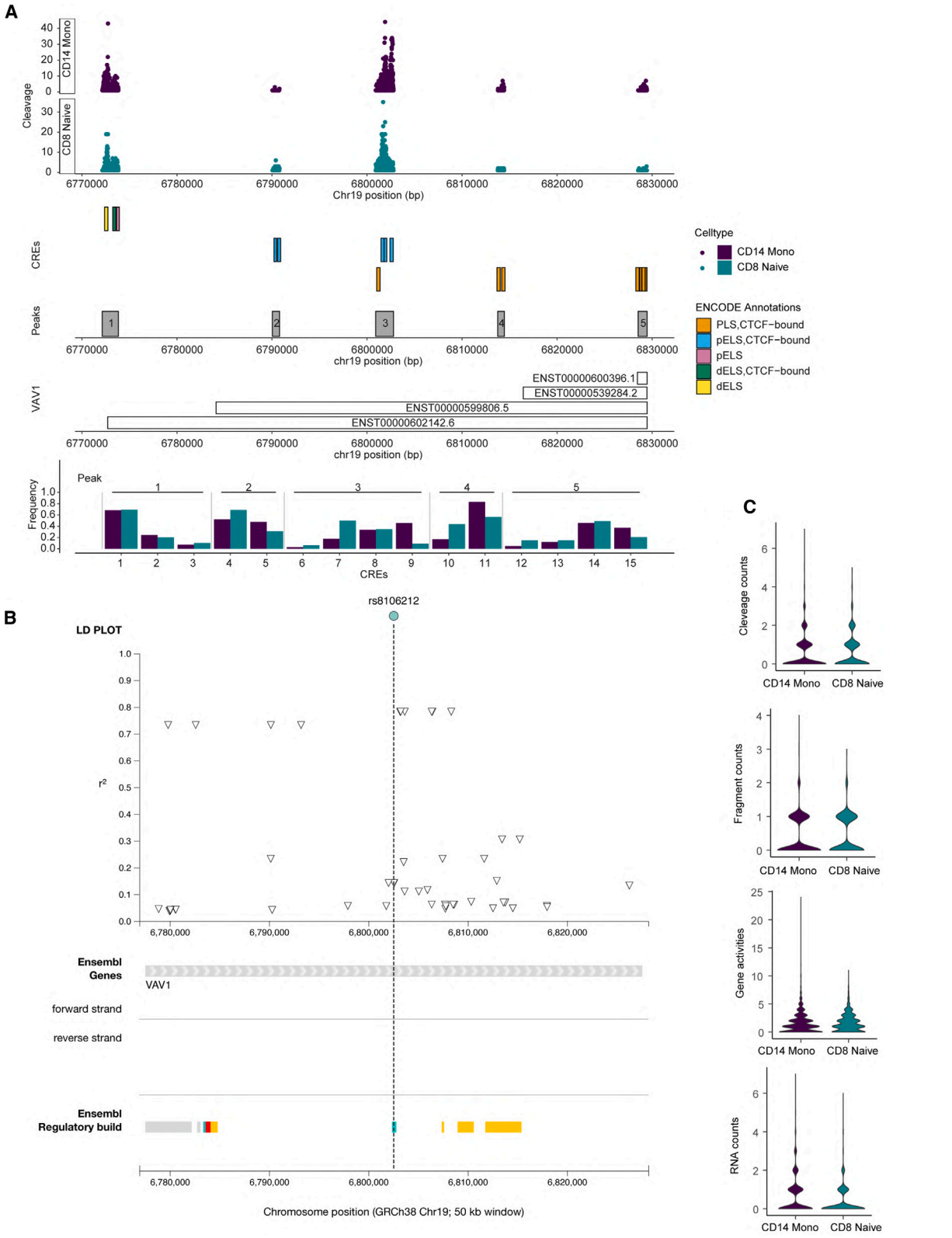
(E) Pie charts depicting the distribution of Tn5 cleavage sites across MACS2 peaks.

(F) Distribution of Tn5 cleavage site-to-peak ratios and the proportion of CRE-associated Tn5 cleavages across MACS2 peaks.

(G) An illustration of a peak near the *SFT2D2* gene exhibiting distinct regulatory patterns between CD14$^+$ and CD8$^+$ cells.

(H) Violin plots of fragment counts, gene activities, and RNA UMI counts across cells in *SFT2D2*.

MACS2 peaks identified in the PBMC multiome dataset, using the paired scRNA-seq data to validate our findings. We highlight three representative examples to show how the use of CRE annotations allowed us to capture regulatory events that we would have missed had we relied solely on peak-level summaries and gene activity scores.



(legend on next page)

First, we identified a peak on chromosome 1 containing six annotated CREs, covering approximately 2 kb and overlapping the promoter and part of the coding region of *SFT2D2* (Figure 1G). We found that $CD14^+$ monocytes and $CD8^+$ naive T cells exhibit distinct cell-type-specific CRE usage patterns. CRE5 and CRE6, which are CTCF-bound enhancers unique to $CD14^+$ monocytes, exhibit markedly lower activity in $CD8^+$ naive cells, which instead show robust CRE2 activity. scRNA-seq reads from the same cell types provide further evidence supporting differential expression of *SFT2D2* between $CD14^+$ monocytes and $CD8^+$ naive cells. This suggests that CRE5 and CRE6 may enhance transcription in $CD14^+$ monocytes (Figures 1H and S6). Notably, $CD14^+$ monocytes and $CD8^+$ naive cells have similar mean fragment counts within this peak, and their *SFT2D2* gene activity scores are comparable. In another region on chromosome 1, we identified six CREs within a peak spanning 2,274 nt and overlapping the gene *STK40* (Figure S7). Here, CRE6 shows increased accessibility in $CD14^+$ monocytes, reflected by a surge in Tn5 cleavage events. This implicates CRE6 as a leading factor driving the expression of *STK40* in $CD14^+$ monocytes. While we noted differential gene activity scores for *STK40* between $CD14^+$ and $CD8^+$ cells, gene activity alone failed to pinpoint the specific regulatory elements involved. These examples (Figure S8) suggest that metrics based solely on scATAC-seq peaks may be insufficient to capture cell-type-specific regulatory differences.

Finally, we highlight a region on chromosome 19 surrounding the gene *VAV1*, which has five MACS2 peaks in its vicinity. ENCODE annotations document at least four transcripts and 15 CREs in this region, with each peak containing two to five CREs (Figure 2A). Notably, within the third and fourth peaks, we observe a distinct pattern including CRE7, -9, -10, and -11. In particular, CRE9 exhibits high accessibility, suggesting it functions as an enhancer specific to $CD14^+$ monocytes. Examining this region more closely, we uncovered a few genome-wide association study (GWAS) SNPs associated with platelet counts, platelet volume, and platelet distribution width, all with significant p values (Figure 2B; Table S1).[20] When activated, platelets bind to monocytes, forming monocyte-platelet aggregates (MPAs) that alter the function and phenotype of the monocytes.[21] This interaction between platelets and monocytes provides a critical link between inflammation and thrombosis. Our analysis shows that CRE9 is highly accessible in $CD14^+$ monocytes but not in other PBMC cell types, indicating that its specific regulatory role in monocytes is to influence platelet phenotypes. As before, we find that analyses focusing on peak-level fragment counts, read counts, or gene activity scores will overlook nuanced details (Figure 2C).

To conclude, interpreting accessibility data solely at the peak or gene level risks missing essential insights into CREs, similar to grains of sand slipping through a sieve. In contrast, utilizing ENCODE CRE annotations enhances our interpretation of scATAC-seq signals. We have formalized this approach into our newly developed CREscendo framework.

Overview of the CREscendo framework

To recover signals that are missed in peak-based analyses, we developed the CREscendo framework, which leverages regulatory annotations and Tn5 cleavage frequencies to capture differential CRE usage within peaks between cell types. Focusing on peaks that contain more than one CRE, CREscendo uses CRE annotations from ENCODE to dissect each peak into distinct regions (Figure 3A). For peaks with k overlapping CREs, we partition the peak into k + 1 segments—one for each CRE (CRE_1, CRE_2, …, CRE_k)—plus an additional segment accounting for base pairs within the peak that are not covered by any CRE. We identify Tn5 cleavage sites using the start- and endpoints of each segment and count the cleavage events for each segment across different cell types. For each peak, we construct a contingency table using the cleavage counts across segments between cell types. CREscendo employs a chi-squared test to assess statistically significant differences in cleavage frequencies, indicating differential CRE usage. Upon obtaining test statistics for all peaks, we adjust for multiple comparisons using the false discovery rate (FDR) method. Furthermore, we refine our analysis by quantifying the contribution of each CRE through the breakdown of the overall chi-squared statistic into partial chi-squared values for each CRE segment.

CREscendo recovers differential CRE usage that peak-based analyses fail to capture, improving interpretability of results

We first applied CREscendo to analyze MACS2 peaks from $CD14^+$ monocytes and $CD8^+$ naive cells (PBMC 10× Multiome data) and compared our results to those from DA analysis using Signac. Out of 53,951 MACS2 peaks encompassing multiple CREs, 13,803 exhibited differential CRE usage patterns that were not identified by Signac. Among these, we identified regions demonstrating strong cell-type-specific regulatory patterns that would be overlooked in current peak-based analyses due to minimal fold changes across the entire peak (Figure 3B). Notable genes that show strong CRE-level signals include *CD55*, *EEF2K*, *RAC2*, and *PDCD4*, among others.

Even when DA peaks are identified by Signac, they offer limited interpretative value, as accessibility over an entire peak can be vague. For example, we identified a peak near *CD248* (also known as *TEM1*) containing four CREs, with cell-type-specific regulatory patterns demarcating CRE3 and CRE4. This peak is ranked low by Signac, as differences in gene activities and fragment counts are minimal. Yet, with CREscendo, this peak is prioritized as significant (chi-squared statistic of 3,353.924, $p < 1e{-}16$, Figures 3C and S9), indicating striking differences in CRE usage between the cell types. CRE3 is highly active in $CD8^+$ naive cells, whereas CRE4 is highly active in $CD14^+$ monocytes. ENCODE annotations indicate that CRE3 might be an enhancer specific to $CD8^+$ naive T cells. Our observations are supported by findings from bulk data collected from related cell types, where both CRE3 and CRE4 display cell type

Figure 2. Detailed view of peaks near the *VAV1* gene

(A) An illustration of peaks near the *VAV1* gene exhibiting distinct regulatory patterns between $CD14^+$ and $CD8^+$ cells.

(B) Linkage disequilibrium plot for SNP rs8106212.

(C) Violin plots of fragment counts, gene activities, and RNA UMI counts across cells in *VAV1*.

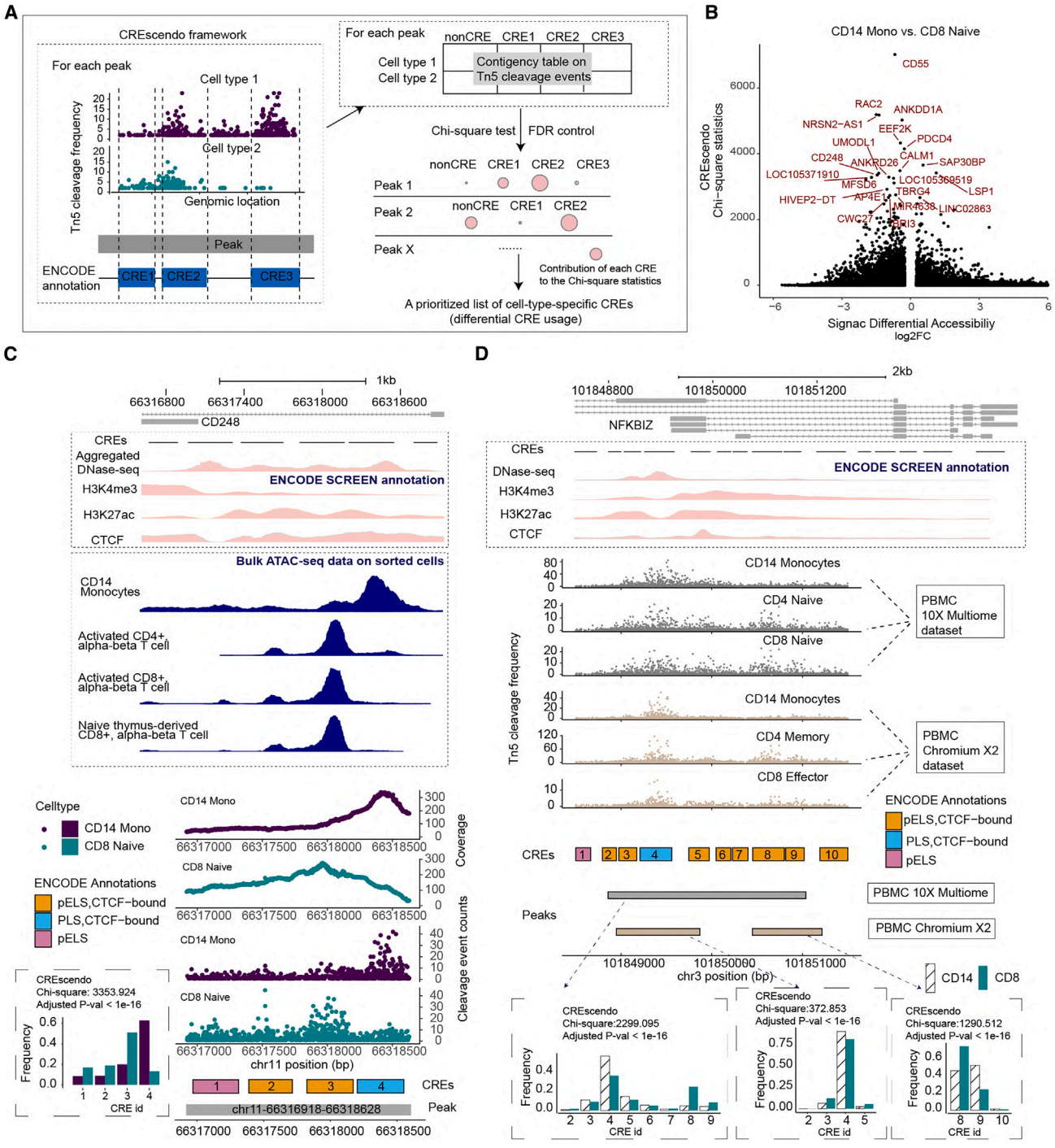


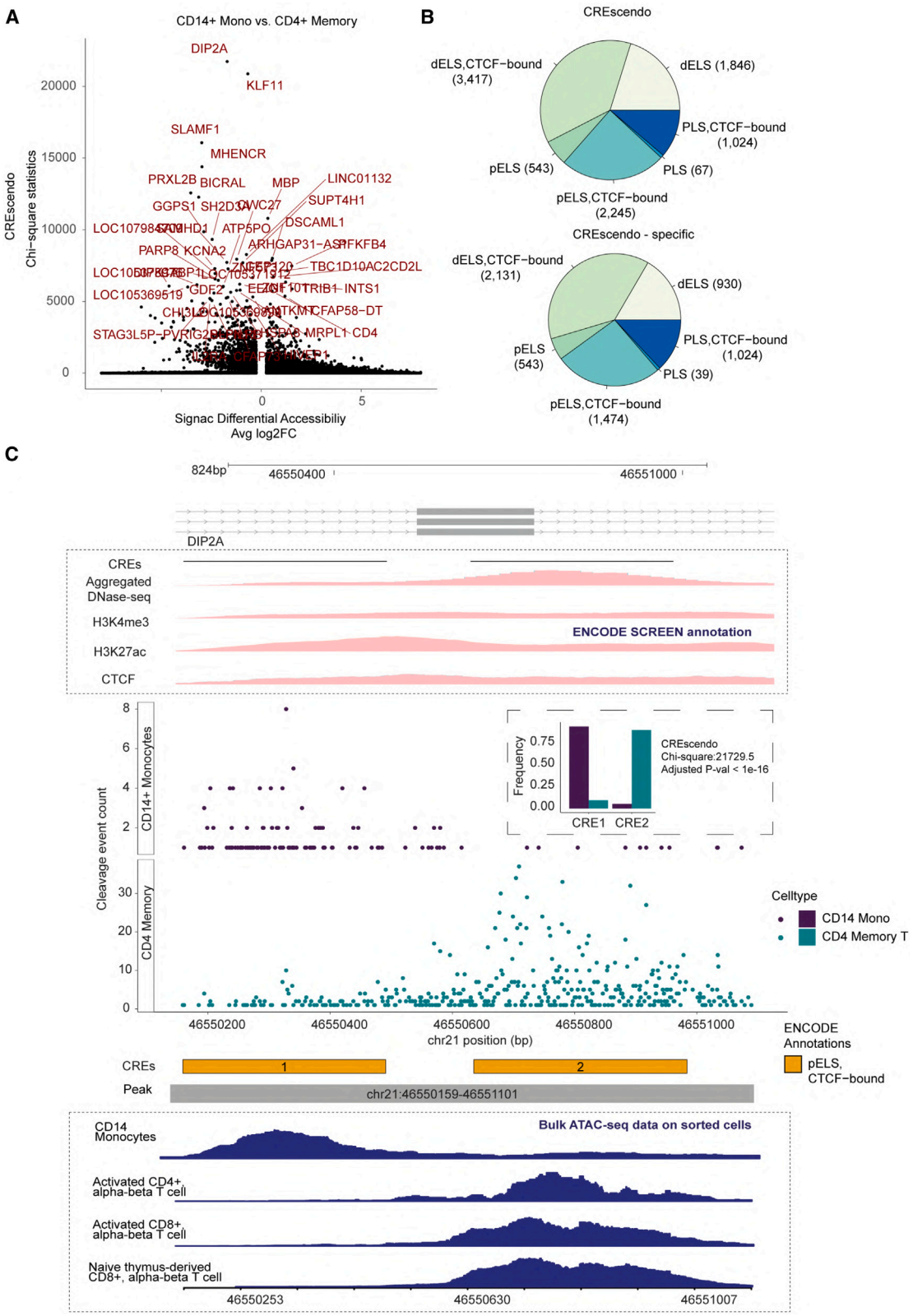
Figure 3. The CREscendo framework can uncover distinct regulatory patterns within scATAC-seq peaks

(A) Schematic representation of the CREscendo analysis highlighting differential usage of CREs.

(B) Scatterplot comparing chi-squared statistics from the CREscendo analysis with log2 fold change from Signac's DA analysis. Peaks with chi-squared statistics greater than 2,000 are highlighted in red text.

(C) Detailed view of a peak near the *CD248* gene compared with ENCODE data annotations.

(D) Detailed view of a peak near the *NFKBIZ* gene compared with data from the PBMC Chromium X2 dataset.



(legend on next page)

specificity (Figure 3C). Specifically, the high activity of CRE3 is recapitulated in activated CD4$^+$ and CD8$^+$ αβ T cells and naive αβ T cells derived from the thymus, whereas high activity of CRE4 is observed in sorted CD14$^+$ monocytes. The function of CD248 is also consistent with these findings, as it encodes endosialin, a stromal cell antigen that is expressed on naive human CD8$^+$ T cells and regulates proliferation.[22]

Next, we extended differential CRE usage analysis to analyze CD14$^+$ monocytes and CD4$^+$ memory T cells in the 10× Chromium X2 dataset. Another pertinent issue with peak-based analysis is the lack of portability and comparability of peaks and results across different samples and studies. For instance, we identified a single peak near *NFKBIZ* (Figure 3D) in the multiome dataset that appears as two distinct peaks in the Chromium X2 dataset. Each of these peaks encompasses more than one annotated cCREs. All peaks show significant differential CRE usage between CD14$^+$ and CD8$^+$ cells in CREscendo's analysis; however, as differential CRE usage is peak dependent, the differences in peak annotation across datasets pose a challenge for interpretation.

We further compared our results to those from DA analysis using Signac. We identified regions demonstrating strong cell-type-specific regulatory patterns that would be overlooked by Signac due to minimal fold changes across the entire peak (Figures 4A and S10–S12). Notable genes that show strong signals at the CRE level include *DIP2A*, *KLF11*, *SLAMF1*, and others. From a total of 83,506 Cell Ranger peaks that encompass multiple CREs, 4,053 showed strong differential CRE usage patterns. Of these, 3,354 peaks are located near coding genes, with 39.8% of them not detected by Signac. Peaks unique to CREscendo that exhibit differential CRE usage affect a total of 5,604 CREs. This indicates potential information leakage during gene-level DA analysis.

Taking *DIP2A* as an example, we identified a peak with cell-type-specific regulatory patterns demarcating CRE1 and CRE2. In CREscendo, this event is prioritized as significant (chi-squared statistics of 21,729.5, $p < 1e{-}16$), indicating striking differences in CRE usage between CD14$^+$ monocytes and CD4$^+$ memory T cells. Within this peak, CRE2 shows high activity in CD4$^+$ memory T cells but is nearly inactive in CD14$^+$ monocytes. According to ENCODE annotations, CRE3 may be an enhancer specific to CD4$^+$ memory T cells. In contrast, Signac does not prioritize this peak, as the differences in gene activities or fragment counts are not substantial. Our results are consistent with bulk data from alternative assays of related cell types (Figure 4C). The high activity of CRE2 is recapitulated in activated CD4$^+$ or CD8$^+$ αβ T cells or in naive αβ T cells from the thymus but not in sorted CD14$^+$ monocytes, underscoring its cell-type specificity.

Finally, we conducted a differential CRE usage analysis to compare L2/3 IT cells (excitatory pyramidal neurons in cortical layer 2/3) and oligodendrocytes in the mouse cortex using the 10× Chromium X2 dataset. 37,048 peaks containing more than one CRE were used for the test. We then compared these results to those obtained from DA analysis using Signac. Our analysis identified cell-type-specific regulatory regions that were not detected by Signac due to minimal fold changes across the entire peak (Figure S13). Notable genes showing strong CRE-level signals in this comparison included *Adcy5*, *Rgs3*, *Sema6d*, and others. Within Cell Ranger peaks that span multiple CREs, 1,171 regions demonstrated significant differential CRE usage. Of these, 25.8% were not detected by Signac (Signac identified 25,294 DA peaks from 40,470 tested). Importantly, peaks unique to CREscendo revealed differential CRE usage, affecting a total of 489 CREs.

We further compared L2/3 IT cells with astrocytes and identified additional genes with strong CRE-level signals, such as *Sybu*, *Tox2*, and *Nim1k*. In this comparison, 1,371 Cell Ranger peaks showed significant differential CRE usage, with 18.3% of these not detected by Signac (Signac reported 33,343 DA peaks from 43,701 tested; Figure S14). Peaks unique to CREscendo in this analysis affected a total of 359 CREs. These results highlight the possibility of information loss in gene-level DA analysis when differential regulatory activity occurs within specific regions of broader peaks.

Among the top 10 CREs showing differential usage in both comparisons, the gene *Agt* (angiotensinogen) emerged as a shared target, with differential activity spanning its promoter region. CREscendo prioritized this event as significant, with chi-squared statistics of 1,698.613 (L2/3 IT vs. oligodendrocytes, $p < 1e{-}16$; Figure 5) and 1,731.345 (L2/3 IT vs. astrocytes; Figure S15). In contrast, *Agt* ranked 2,641 out of 25,294 significant DA peaks in the L2/3 IT vs. oligodendrocyte comparison and 3,731 out of 33,343 in the L2/3 IT vs. astrocyte comparison under Signac. This highlights striking differences in CRE usage between cell types and emphasizes the power of CREscendo in identifying regulatory patterns. Based on the bulk ATAC-seq and chromatin immunoprecipitation (ChIP)-seq data from ENCODE, CRE1 and CRE2 appear to function as distinct CREs. This is supported by their different activity across brain regions and distinct binding profiles of transcription factors, as indicated by the unique signal patterns observed in the data. L2/3 IT neurons likely depend on specific enhancers of *Agt* for roles such as neurotransmitter modulation, synaptic plasticity, or local production of angiotensin II to support neuronal activity and connectivity. Oligodendrocytes, by contrast, may utilize distinct *Agt* enhancers, reflecting their role in myelination and possibly in mitigating oxidative stress or inflammation in response to neuronal signals. Finally, astrocytes are known as the primary producers of angiotensinogen in the CNS. Differential or enhanced *Agt* activity in astrocytes is critical for the local production of angiotensin peptides, which regulate blood-brain barrier integrity, cerebral blood flow, and neuronal excitability.

Figure 4. Applying the CREscendo framework to differential CRE usage in CD14$^+$ monocytes and CD4$^+$ memory T cells in the 10× Chromium X2 dataset

(A) Scatterplot comparing chi-squared statistics from the CREscendo analysis with log2 fold change from Signac's DA analysis. Peaks with chi-squared statistics greater than 2,000 are highlighted in red text.

(B) Pie chart of CRE annotations for CREs in peaks with differential CRE usage identified by CREscendo (top) and uniquely by CREscendo (bottom).

(C) Detailed view of a peak near the *DIP2A* gene compared with ENCODE data annotations.

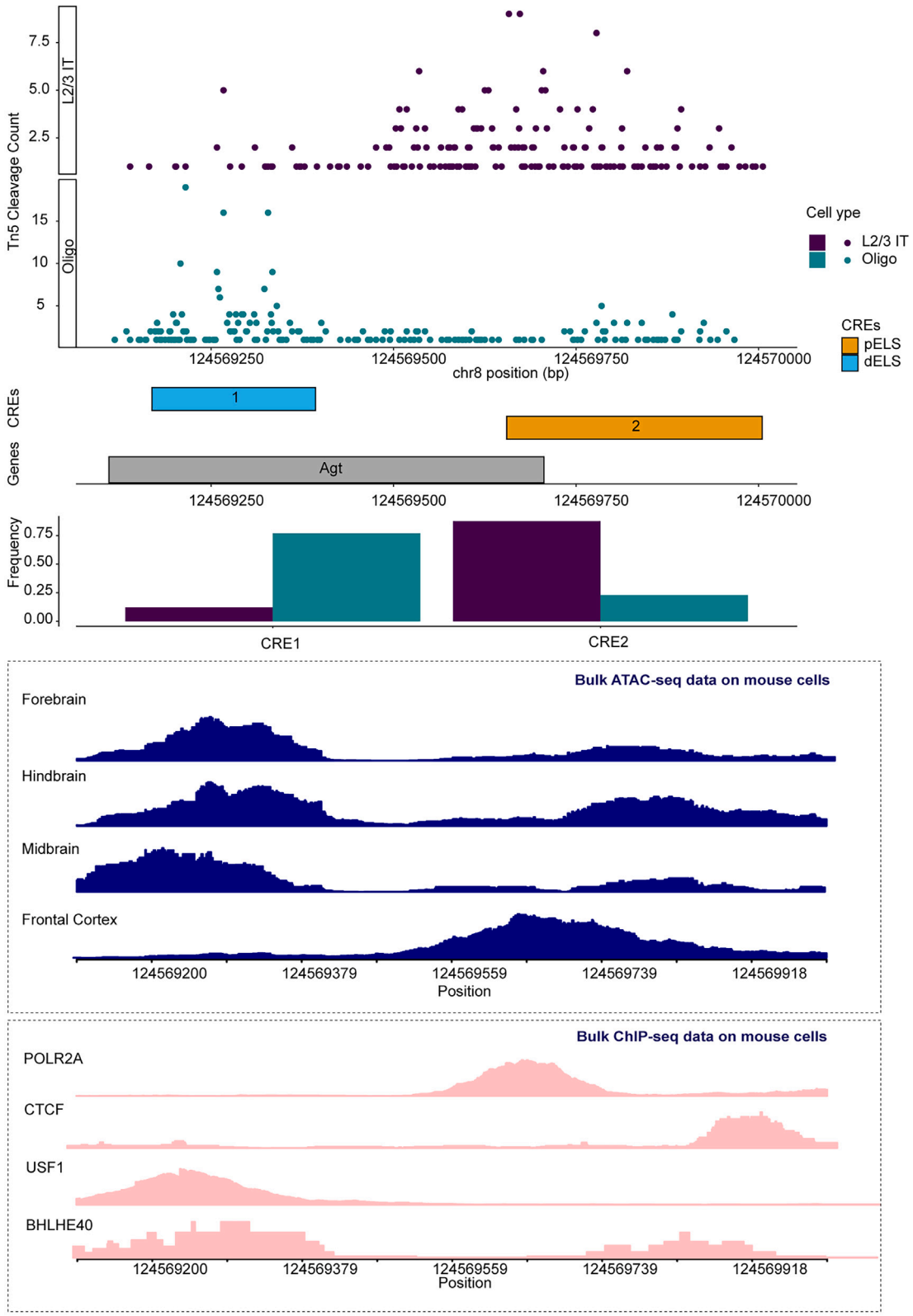


Figure 5. Detailed view of a peak near the *Agt* gene from differential CRE usage in L2/3 IT and oligodendrocytes from the adult mouse cortex Chromium X2 dataset

DISCUSSION

Our analysis underscores key limitations inherent in peak-based approaches for scATAC-seq analysis. The genomic landscape depicted by peak-based analyses is a landscape that is occluded by fog, where individual structures blur together into indistinct shapes. In other words, peak-based approaches enable the identification of large-scale regulatory features but often fail to resolve cell-type-specific CREs, which are crucial for understanding fine-grained gene regulation. This lack of resolution not only impacts the sensitivity of downstream analyses but also raises concerns about the rigor and quality control of peak definitions across different studies. The variability in peak calling can lead to results that lack portability and are challenging to interpret, compromising the reproducibility of research findings.

To address these challenges, we propose the adoption of the CREscendo framework, which utilizes a standardized reference of annotated CREs to quantify scATAC-seq data. Our findings indicate that the CREscendo framework can clarify the genomic landscape, reducing the ambiguities caused by arbitrary peak definitions and revealing signals that are often missed in peak-based analyses. However, CREscendo is no panacea, as it still operates within peaks, which means it is not entirely free from the constraints imposed by peak-based methods.

Despite these constraints, our research strongly supports the transition from a peak-based approach to scATAC-seq quantification to one grounded in a standardized reference of annotated CREs. This shift is not just a technical adjustment but a philosophical one, pushing the field toward a more systematic study of CREs with defined coordinates, akin to how genes are studied. A starting point will be to use the CRE-level summary as the metric to quantify scATAC-seq signals and perform downstream analyses, including clustering and DA analysis. Looking forward, adopting a peak-free approach tailored to well-annotated genomes, like that of humans, is a paradigm shift toward precision and reproducibility in scATAC-seq analyses. By focusing directly on annotated CREs, researchers can achieve more interpretable, reproducible, and biologically meaningful insights, paving the way for a systematic exploration of the genome's regulatory code. While transitioning to a CRE-centered approach offers clear advantages, it is not without its challenges. The field must address issues like incomplete annotations, computational overhaul, and resistance to change. Simultaneously, peak-based analyses retain their value for quality control, exploratory discovery, and compatibility with existing tools. A gradual, hybrid transition that combines the strengths of both approaches may offer the most practical path forward, ensuring that researchers can leverage the benefits of CRE-based methods while maintaining the robustness and flexibility of peak-based workflows.

Limitations of the study

While standardization improves reproducibility, it might also reduce flexibility in exploring new regulatory elements that fall outside of the current CRE annotations. The field may miss novel insights that require a more exploratory approach, such as discovering non-canonical or emerging CREs. Moreover, the success of the transition to CRE-based scATAC-seq analysis is heavily dependent on the quality and comprehensiveness of existing databases like ENCODE. Any limitations or biases in these databases will directly impact the effectiveness of the analysis, potentially leading to biased interpretations.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Mengjie Chen (mengjiechen@uchicago.edu).

Materials availability

This study did not generate unique reagents.

Data and code availability

We provide an R package that implements proposed methods discussed in this study. The CREscendo package is available from GitHub (https://github.com/ChenMengjie/CREscendo). In addition, the R source code to reproduce all data analysis in the study is available under the same GitHub directory as the tutorials and on Zenodo at https://doi.org/10.5281/zenodo.14788274.

ACKNOWLEDGMENTS

The author thanks Natalia Gonzales, Sarah Sumner, and Christian Jones for their comments and edits, which significantly improved the manuscript. The author thanks Lei Zheng for assistance extracting bulk ATAC-seq data. The author thanks Yan Li for assistance running the peakVI tutorial. The work was supported by National Institutes of Health grants R01 GM126553, R01 HG011883, and HG012927 to M.C.

AUTHOR CONTRIBUTIONS

M.C. conceived this work, developed the methods, performed the analyses, and wrote the paper.

DECLARATION OF INTERESTS

The author declares no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Data and preprocessing
  - CRE annotation
  - CRE scendo implementation
  - Differential accessibility analysis
  - Bulk ATAC-seq data

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2025.100806.

Received: October 22, 2024
Revised: January 4, 2025
Accepted: February 11, 2025
Published: March 5, 2025

REFERENCES

1. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.
2. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat. Methods *18*, 1333–1341.
3. Martens, L.D., Fischer, D.S., Yépez, V.A., Theis, F.J., and Gagneur, J. (2024). Modeling fragment counts improves single-cell ATAC-seq analysis. Nat. Methods *21*, 28–31.
4. Teo, A.Y.Y., Squair, J.W., Courtine, G., and Skinnider, M.A. (2024). Best practices for differential accessibility analysis in single-cell epigenomics. Nat. Commun. *15*, 8805.
5. Yan, F., Powell, D.R., Curtis, D.J., and Wong, N.C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol. *21*, 1–16.
6. Yuan, Q., and Duren, Z. (2024). Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. Nat. Biotechnol., 1–11.
7. Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., and Aerts, S. (2023). SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. Nat. Methods *20*, 1355–1367.
8. Ashuach, T., Reidenbach, D.A., Gayoso, A., and Yosef, N. (2022). PeakVI: A deep generative model for single-cell chromatin accessibility analysis. Cell Rep. Methods *2*, 100182.
9. Chen, H., Lareau, C., Andreani, T., Vinyard, M.E., Garcia, S.P., Clement, K., Andrade-Navarro, M.A., Buenrostro, J.D., and Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. *20*, 1–25.
10. Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., et al. (2022). A Python library for probabilistic analysis of single-cell omics data. Nat. Biotechnol. *40*, 163–166.
11. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. *53*, 403–411.
12. Ji, Z., Zhou, W., Hou, W., and Ji, H. (2020). Single-cell ATAC-seq signal extraction and enhancement with SCATE. Genome Biol. *21*, 1–36.
13. Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978.
14. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 1–5.
15. Yu, W., Uzun, Y., Zhu, Q., Chen, C., and Tan, K. (2020). scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. Genome Biol. *21*, 1–17.
16. Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. *37*, 925–936.
17. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, 1–9.
18. Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. *20*, 207–220.
19. Stuart, T., Srivastava, A., Lareau, C., and Satija, R. (2020). Multimodal single-cell chromatin analysis with Signac. Preprint at bioRxiv. https://doi.org/10.1101/2020.11.09.373613.
20. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.
21. Larsen, E., Celi, A., Gilbert, G.E., Furie, B.C., Erban, J.K., Bonfanti, R., Wagner, D.D., and Furie, B. (1989). PADGEM protein: a receptor that mediates the interaction of activated platelets with neutrophils and monocytes. Cell *59*, 305–312.
22. Valdez, Y., Maia, M., and Conway, E.M. (2012). CD248: reviewing its role in health and disease. Curr. Drug Targets *13*, 432–439.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
10X Genomics PBMC Multiome Dataset	This paper	https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0
10X Genomics PBMC Chromium X2 ATAC-seq Dataset		https://www.10xgenomics.com/datasets/10k-human-pbmcs-atac-v2-chromium-x-2-standard
10X Genomics Adult Mouse Cortex Chromium X2 ATAC-seq Dataset		https://www.10xgenomics.com/datasets/8k-adult-mouse-cortex-cells-atac-v2-chromium-controller-2-standard
Signac PBMC scRNA-seq Dataset		https://stuartlab.org/signac/articles/pbmc_vignette
ENCODE Registry of Candidate CREs v.3		https://screen.encodeproject.org/index/cversions
Bulk ATAC-seq Bigwig Tracks		GEO: GSM5513655
Software and algorithms		
CREscendo	This manuscript	https://github.com/ChenMengjie/CREscendo
CellRanger		https://www.10xgenomics.com/support/software/cell-ranger/latest
Signac		https://stuartlab.org/signac/
MACS2		https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html
MACS3		https://github.com/macs3-project/MACS

METHOD DETAILS

Data and preprocessing

We downloaded the PBMC_multiome dataset, which includes peak files and fragment files, from 10x Genomics (https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0). The multiome platform produces data that include both ATAC and RNA profiles from the same cells. Each cell's data are linked by a unique cell barcode. The dataset is estimated to contain 11,898 cells. We conducted quality control using Signac (v1.11.0), applying the following criteria: nCount_ATAC less than 100,000, nCount_RNA less than 25,000, nCount_ATAC greater than 1,800, nCount_RNA greater than 1,000, nucleosome signal less than 2, and TSS enrichment greater than 1. This quality control process resulted in 11,070 cells with 108,377 peaks identified by Cellranger (v2.0.0). Following this, we employed MACS2 (as implemented in Signac) for peak calling on cells post-quality control, which yielded a list of 131,364 peaks. We applied Signac for clustering and cell type annotations. We performed most of our cell type-specific analysis in the main text using $CD14^+$ Monocytes (3,089 cells) and $CD8^+$ Naive T cells (1,502 cells). The analysis of other cell types is presented in supplementary materials.

We downloaded the PBMC Chromium X2 scATAC-seq dataset, which includes peak and fragment files, from 10x Genomics (https://www.10xgenomics.com/datasets/10k-human-pbmcs-atac-v2-chromium-x-2-standard). The dataset initially contained 10,273 PBMC nuclei. We performed quality control using Signac (v1.11.0), using the following criteria: nCount_peaks between 3,000 and 30,000, pct_reads_in_peaks greater than 15%, and nucleosome_signal less than 4. This process yielded 6,728 cells across 164,487 peaks, as analyzed by Cellranger (v2.1.0). Additionally, we used Signac to annotate cell types by aligning with a pre-processed PBMC scRNA-seq dataset available from (https://stuartlab.org/signac/articles/pbmc_vignette). The annotated cell types included pre-B cells (352), $CD14^+$ Monocytes (1,224), Double-negative T cells (135), $CD4^+$ Memory T cells (4,108), $CD8^+$ Effector T cells (430), and NK dim cells (479). These cell types differ from those identified in the PBMC_multiome dataset. To benchmark with results from PBMC_multiome, we performed cell type-specific analysis in the main text using $CD14^+$ Monocytes vs. $CD4^+$ Memory T cells.

We downloaded the Adult Mouse Cortex Chromium X2 scATAC-seq dataset, which includes peak and fragment files, from 10x Genomics (https://www.10xgenomics.com/datasets/8k-adult-mouse-cortex-cells-atac-v2-chromium-controller-2-standard). The dataset initially contained 7,729 cortex nuclei. We performed quality control using Signac (v1.11.0), using the following criteria: nCount_peaks between 3,000 and 100,000, pct_reads_in_peaks greater than 40%, nucleosome_signal less than 4, blacklist_ratio less than 0.025 and TSS.enrichment >2. This process yielded 5,310 cells across 199,032 peaks, as called from Cellranger (v2.1.0). Additionally, we used Signac to annotate cell types by aligning with a pre-processed PBMC scRNA-seq dataset available

from (https://stuartlab.org/signac/articles/mouse_brain_vignette). The annotated cell types included L2/3 IT (1066), L4 (884), Oligodendrocytes (811), Astrocytes (575), L6 CT (485) and others. We performed cell type-specific analysis using L2/3 IT, Oligodendrocytes and Astrocytes.

CRE annotation

For the annotation of *cis*-regulatory elements (CREs), we utilized resources from ENCODE, specifically the Registry of candidate CREs Version 3 from SCREEN (https://screen.encodeproject.org/index/cversions). This resource contains 1,063,878 candidate CREs for human and 368,121 candidate CREs for mouse.

CRE scendo implementation

We focus on peaks that encompass more than one CRE. Tn5 cleavage sites are identified by using the start and end sites of each segment from fragment files. We annotate each peak using ENCODE-defined CREs. For peaks with k overlapping CREs, we divide the peak into k+1 segments (CRE_1, CRE_2, …, CRE_k) plus an additional segment for base pairs within the peak not covered by any CREs. For each cell type, we construct a frequency vector that represents the number of cleavage sites falling within each segment. We perform a chi-square test for each peak with degrees of freedom equal to k. In the chi-square tests for comparison, with two cell types A and B, we conduct the test first using A as the reference population, and then repeat using B. The final *p*-value is determined as the maximum (less significant) from these two tests. After generating test statistics for all peaks, we apply false discovery rate (FDR) adjustments to correct for multiple comparisons. We further quantified the contribution of CRE by breaking down the overall chi-square statistic into partial chi-square statistics from each CRE. The above procedure is implemented in the CREscendo package.

To refine our results from the chi-squared test, we implement the Filter() function. This function is designed to extract and return high-confidence results from the summary component of the chi-square test, enabling prioritized downstream analysis. The Filter() function accepts the following parameters: - x: A CREscendoTested object containing the results from CRE association tests. - fdr_cutoff: A numeric value specifying the FDR adjusted *p*-value threshold for filtering the results. The default is 0.05. - coverage_cutoff: A numeric value specifying the cutoff of the number of fragments across all cells within the same type for each peak region for filtering the results. The default is 100. - abs_diff_cutoff: A numeric value specifying the cutoff of the absolute difference in the CRE frequencies between two cell types. The default is 0.2. By applying these filtering criteria, we can focus our analysis on the most significant relevant CRE associations. The analysis presented in the manuscript applied the default parameters.

We further implement a Visualize() function to visualize cleavage sites of different cell types for a specific peak or peaks from a specific gene within the CREscendo test object. This function facilitates the extraction and visualization of relevant data, including cleavage sites, CRE annotations, and CRE frequencies within tested cell types. All related figures presented in the manuscript were generated using the Visualize() function. A diagram that describes the workflow of CREscendo can be found in Figure S16.

Differential accessibility analysis

We conducted differential accessibility (DA) analysis using the logistic regression (LR) implementation provided in Signac. This approach introduces the total number of fragments as a latent variable to compensate for variation in sequencing depth. For the PBMC 10X multiome dataset, we performed DA tests between different cell types. In our comparison between $CD14^+$ Monocytes and $CD8^+$ Naive T cells, we input all 131,364 MACS2 peaks. The default analysis using the FindMarkers(…, test.use = 'LR', latent.vars = 'nCount_peaks') function identified 24,534 peaks with *p*-values less than 0.05. After adjusting for multiple comparisons, 22,699 peaks remained significant. Of these, 20,695 peaks overlapped with peaks used in the CREscendo analysis, indicating that these regions contain more than one CRE.

For the PBMC 10X Chromium X2 dataset, we performed DA between $CD14^+$ Monocytes and $CD4^+$ Memory T cells, using all 164,487 Cellranger peaks as input. The default analysis using the FindMarkers(…, test.use = 'LR', latent.vars = 'nCount_peaks') function identified 29,110 peaks with *p*-values less than 0.05. After adjusting for multiple comparisons, 25,447 peaks remained significant. Of these, 24,954 peaks overlapped with peaks used in the CREscendo analysis, indicating that these regions contain more than one CRE.

For the adult mouse cortex Chromium X2 scATAC-seq dataset, we conducted DA analysis between L2/3 IT and oligodendrocyte cells, using all 199,032 Cellranger peaks as input. The analysis, performed with the FindMarkers function (test.use = 'LR', latent.vars = 'nCount_peaks), initially identified 40,470 peaks with *p*-values less than 0.05, of which 25,294 remained significant after adjusting for multiple comparisons. Among these, 8,125 peaks overlapped with regions used in the CREscendo analysis, indicating that these regions harbor more than one CRE.

Using the same settings, we also analyzed DA between L2/3 IT and astrocyte cells. This analysis identified 43,701 peaks with *p*-values below 0.05, with 33,343 retaining significance after multiple testing correction. Of these, 11,658 peaks overlapped with the CREscendo regions, indicating the presence of multiple CREs in these loci.

Bulk ATAC-seq data

We first reviewed the ENCODE data available for related T cell types. Bulk ATAC-seq data are available for activated $CD4^+$ and $CD8^+$ $\alpha\beta$ T cells, as well as Naive thymus-derived $CD8^+$ $\alpha\beta$ T cells. We extracted relevant bigwig tracks from the ENCODE region browser.

Bigwig tracks of bulk ATAC-seq on $CD14^+$ Monocytes isolated from whole blood buffy coats were download from GEO (GSM5513655).

We reviewed the ENCODE data available for mouse studies. Unfortunately, there is no matched tissue or cell lines for the cell types used in the analysis. We instead downloaded the most relevant datasets including bulk ATAC-seq data from forebrain, hindbrain and midbrain tissue from postnatal 0 days mouse from strain B6NTac; B6NCrlLap as well bulk ATAC-seq data from adult frontal cortex tissue from strain B6NCrl. For bulk ChIP-seq data, we extracted bigwig tracks from the ENCODE project for CTCF and POLR2A from strain B6NCrl, and for USF1 and BHLHE40 from cell line MEL.