Efficient pre-processing of Single-cell ATAC-seq data

Fan Gao[1,2], Lior Pachter[1,3*]

[1]Division of Biology and Biological Engineering; [2]Caltech Bioinformatics Resource Center; [3]Department of Computing & Mathematical Sciences

California Institute of Technology, Pasadena, CA, USA

*To whom correspondence should be addressed.

ABSTRACT

The primary tool currently used to pre-process 10X Chromium single-cell ATAC-seq data is Cell Ranger, which can take very long to run on standard datasets. To facilitate rapid pre-processing that enables reproducible workflows, we present a suite of tools called scATAK for pre-processing single-cell ATAC-seq data that is 18 times faster than Cell Ranger on human samples, and that uses 33% less RAM when 8 CPU threads are used. Our tool can also calculate chromatin interaction potential matrices and generate open chromatin signals and interaction traces for cell groups. We demonstrate the utility of scATAK in an exploration of the chromatin regulatory landscape of a healthy adult human brain and show that it can reveal cell-type-specific features.

scATAK is available at https://github.com/pachterlab/scATAK/.

Keywords:

single-cell, ATAC-seq, bioinformatics, open chromatin landscape, chromatin interactome

INTRODUCTION

The development of automated high-throughput single-cell platforms for single-cell ATAC-seq (scATAC-seq) are facilitating highly resolved chromatin accessibility measurements that are valuable in functional genomics studies [1]. The raw data produced in scATAC-seq experiments consists of large numbers of reads, whose pre-processing to identify “peak” regions and counts can pose a formidable challenge. 10X Genomics’ Chromium-based scATAC-seq solution generates data that can be analyzed with companion software called Cell Ranger. While Cell Ranger provides a turnkey solution for labs generating data with 10X’s system, its slow runtime and large memory requirements hinder the development of reproducible workflows for data analysis. In previous work, we have presented a modular and efficient approach to single-cell RNA-seq (scRNA-seq) pre-processing that combines the pseudoalignment program kallisto [2] with a suite of tools called bustools [3]. These tools facilitate the development of highly efficient and modular workflows for scRNA-seq pre-processing that are easy to run using a wrapper called kb [4]. The analysis of single-cell ATAC-seq data requires the mapping of reads to the genome, a more challenging problem than transcriptome alignment due to the size of the genome. Recently, Giansanti et al. [5] developed a pseudoalignment approach for ATAC-seq based on kallisto pseudoalignment of reads to pre-defined DNAseq hypersensitive sites. With their workflow, they were able to produce results in-line with standard results, but much faster with a substantially smaller memory footprint. Motivated by their result, we have incorporated kallisto, bustools, and several other tools into a new scATAC-seq software suite, namely scATAK, that facilitates the processing of scATAC-seq data without the need for pre-defined genome regions.

RESULTS

Overview of scATAK and benchmarking

An overview of scATAK procedure is shown in **Figure 1**. As noted, scATAK is a command-line tool with three modules: *quant*, *track*, and *hic*. Module *quant* runs the following steps for single-cell level quantification: 1) Raw 10X scATAC-seq FASTQ data are processed to add cell barcode sequences from R2 reads to the header lines of R1 and R3 biological reads; 2) Barcode-tagged R1 and R3 FASTQ files for every sample are treated as pseudo-bulk ATAC-seq data for genome alignment using Minimap2 [6], converted to a name sorted BAM alignment file using Sambamba [7], and then subject to peak calling using Genrich [8]; 3) Called peak regions for all samples are merged to generate a list of accessible chromatin regions using bedtools [9] for creating a kallisto index file; 4) With the accessible region index as reference, raw scATAC-seq files are revisited to generate a single-cell region count matrix for every sample using kallisto and bustools; 5) To estimate gene activity for every single cell, we calculate the absolute distance $\boldsymbol{d}$ from the ATAC-seq peak centers to transcription start sites (*TSS*) and associate peak regions to the nearest gene *TSS*, a strategy similar to HOMER [10] peak annotation. Activity score $\boldsymbol{S}$ for gene $\boldsymbol{i}$ is calculated as weighted sum of associated peaks $\boldsymbol{P}$, with $\boldsymbol{S_i} = \Sigma\, \boldsymbol{W_{ij}} \times \boldsymbol{P_j}$, where $\boldsymbol{W}$ is a distance-dependent step function for weight, values from 1 ($\boldsymbol{d} \leq 2$ kb), 0.7 (2 kb < $\boldsymbol{d} \leq 5$ kb), 0.5 (5 kb < $\boldsymbol{d} \leq 10$ kb), 0.25 (10 kb < $\boldsymbol{d} \leq 20$ kb) to 0.03 (20 kb < $\boldsymbol{d} \leq 50$ kb). A distance-dependent weight was originally proposed in the MAESTRO pipeline [11] to better model gene activity. Instead of using computationally expensive exponential decay to calculate $\boldsymbol{W}$, we simply employed a step function to speed up processing. With the accessible region and gene activity count matrices, further analyses can be performed within R or Python notebooks using secondary analyses tools like Seurat [12], snapATAC [13], or chromVAR [14]. For our proof-of-principle analyses, we created R notebooks with DropletUtils [15] and Seurat. After cell clustering and annotation are completed, the scATAK *track* module uses cell barcode — cell group table together with pseudo-bulk ATAC-seq alignment files generated by scATAK *quant* to create cell group bigwig tracks (normalized by the number of cells in the group) for visualization in a genome browser. An additional scATAK *hic* module utilizes a provided bulk HiC [16] or HiChIP [17] interactome map together with a single-cell accessible chromatin region matrix to infer potential chromatin looping events for individual cells and generate group HiC interaction

tracks. Thus group chromatin accessibility and interaction landscapes can be visualized side-by-side.

For benchmarking purposes, we downloaded two 10X scATAC-seq datasets that are hosted on the 10X Genomics dataset website [18] [19]. We processed 224,636,372 raw read pairs from a human PBMC 5k data and 244,056,346 raw read pairs from an adult mouse brain 5k data using both scATAK *quant* and the CellRanger (atac-1.2.0) software. With 2, 4, 8 CPU threads, real run time and memory usage were monitored by snakemake pipeline [20]. As shown in **Table 1**, when PBMC data was processed, scATAK *quant* was roughly 18, 24, 25 times faster than Cellranger with 8, 4, 2 CPU threads employed. For mouse brain data, scATAK was about 15, 18, and 19 times faster when using 8, 4, and 2 CPU threads respectively. With only 2 threads, scATAK *quant* finished PBMC data pre-processing within two and half hours, a reasonable time window for users with limited computational resources to process scATAC-seq data. In contrast, Cellranger took almost 58 hours to process the same data. Also noted from **Table 2**, kallisto bus pseudo-alignment method works well for mapping raw reads to ATAC-seq peak regions, with 45% and 49% pseudo-alignment rates for human and mouse data, respectively. Statistics from bustools showed most of the aligned read pairs (90% for human and 96% for mouse) contain the precise whitelist cell barcodes. With 1-base mismatch barcode error correction method embedded in bustools, 94% and 97% of aligned read pairs remained for single-cell quantification. Inspired by the ultrafast processing speed of scATAK *quant*, we next loaded accessible region count matrices from both scATAK and Cell Ranger to DropletUtils tool to identify cells from empty droplets (FDR ≤ 1e-5, **Figure 2A**). As noted, 3,528 cell barcodes were shared between 3,595 filtered barcodes from scATAK and 3,653 filtered barcodes from Cell Ranger, suggesting a similar data structure of the two matrices. Regions detected in more than 10% of total cells were used for further dimensional reduction and cell clustering. Separate runs of scATAK and Cell Ranger matrices using the default settings of Seurat (with Latent Semantic Indexing to learn the structure of the data [21]) both resulted in 10 cell clusters. The adjusted RAND index for similarity measurement reported a value of 0.915 for cell clustering resulting from scATAK and Cell Ranger. Within each cluster, the vast majority of cells clustered with the scATAK pre-processing were still grouped the same way as the cells clustered after Cell Ranger pre-processing (**Figure 2B**). In other words, cell clusters generated from the two different workflows are highly concordant with each other.

We next loaded gene score information from scATAK *quant* to guide cell type annotation, with *IL7R*, *CD8A* for T cells, *MS4A1* for B cells, *NCR1* for NK cells, *MS4A7* for monocytes, and *ITGAM* for dendritic cells (**Figure 2C**). 10 cell clusters were then merged into 5 groups for different cell types. With chromVAR, we scanned consensus sequences of 386 known human TFs in JASPAR core database (2018 version) and calculated normalized z-scores as a measure for the enrichment of TF motifs at accessible sites of individual cells. With Seurat *Findmarkers* function, signature motifs for different cell clusters were identified (Wilcox test, adjusted p-value < 0.05). Interestingly, MA0824.1_ID4 (**Figure 2D**) showed up as one of the top 20 TF motifs in both c5 (B cell) and c9 (dendritic cell) clusters. This observation is consistent with the regulatory roles of Id proteins in lymphocyte development [22]. Overall, secondary analyses using pre-processed results from the scATAK pipeline revealed expected biological insight from PBMC cells.

Exploration of chromatin accessibility and interaction landscapes in human brain

The brain is a complex organ with highly diversified cell populations. Distinct chromatin landscapes drive cell-type-specific gene expression patterns. Previous large cohort genome-wide association studies (GWAS) unveiled thousands of single nucleotide polymorphisms (SNPs) associated with different neurological disorders, with the majority of SNPs being non-coding variants. Although potential regulatory gene targets of non-coding SNP regions can be identified with high-resolution genome-wide chromatin interactome maps, such analyses do not provide cell-type specificity. As mentioned above, with scATAK we developed a module called *hic* to infer single-cell chromatin looping from bulk chromosome conformation capture (3C) data and scATAC-seq data. Recent technological advances in the 3C field have led to HiChIP [17] – a technology combining chromosome conformation capture with immunoprecipitation- and tagmentation-based library preparation, as a highly sensitive and specific assay to profile chromatin interactions of regulatory chromatin regions. In our analysis, we downloaded a human hippocampal scATAC-seq data together with histone H3K27ac HiChIP data generated from the same brain region of the same individual for integrative analysis [23] (GEO accession numbers

GSM4441823 and GSM4441836). A total of 6,244 cell nuclei were recovered from DropletUtils and 13 cell clusters were generated using Seurat with latent semantic indexing (LSI) to reduce the dimensionality of the scATAC-seq data. Guided by gene activity scores of known brain cell-type-specific marker genes *SLC17A7* (excitatory neurons), *GAD2* (inhibitory neurons), *MAG* (oligodendrocytes), *PDGFRA* (OPC), *GFAP* (astrocytes), and *CX3CR1* (microglia) (**Figure 3A, B**), cells were assigned to one of six major brain cell types. In this scATAC-seq dataset, we found the oligodendrocytes as the major cell population, consistent with Figure 1e of the original paper [23]. Also noted, the complex structure (several clusters) of the open chromatin landscape in excitatory neurons is consistent with multi subtypes of excitatory neurons observed from brain single-cell RNAseq data [24,25], demonstrating chromatin accessibility as a useful molecular marker for sub-clustering of excitatory neurons.

We next asked how genetic variants could explain the susceptibility of hippocampal cells to AD. The scATAK *track* module generated group ATAC signal tracks (normalized by the mapped group read counts) from cell barcode – cell group table and sample pseudo-bulk alignment file. A circos plot (**Figure 3C**) provided a genome-wide view of human GWAS AD risk SNPs [26] (https://ctg.cncr.nl/documents/p1651/AD_sumstats_Jansenetal_2019sept.txt.gz, SNPs with $p<1\times10^{-9}$ included) and ATAC signals in different cell types (signals binned for every 200kb genomic window). AD risk SNPs were further associated with 2kb genomic bins to calculate chromatin accessibility in different cell types. A density plot of astrocytes, microglia, and oligodendrocytes are enriched with subsets of SNP regions that are highly accessible ($\log_{10}$(ATAC-signal + 1)>3, **Figure 3D**). This observation suggests these cell types are vulnerable to AD-associated genetic variation. We next loaded the scATAK *hic* module to subset genomic looping bin pairs (10 kb resolution interaction map (GSM4441836) identified from bulk histone H3K27ac HiChIP data using single-cell chromatin accessibility map already created in scATAK *quant* step. The downloaded map was generated by HiC-Pro [27] to include cis-interactions between 20kb and 2Mb. We further filtered the table and only included bias-corrected significant interactions (Q-Value_Bias <0.05). Assuming open chromatin regions carrying active histone enhancer marks frequently loop together for transcriptional regulation, interacting chromatin pairs (detected in bulk data) that both are accessible regions in individual cells are given a binary potential score for that particular cell. This assumption originated from the observation that the pattern of accessibility variation in *cis* recapitulates chromosome

compartments, linking single-cell accessibility to 3D genome organization, reported by Greenleaf's lab [28]. For N accessible regions in a single cell, $N\times(N-1)/2$ possible combinations will be scanned to find potential looping pairs. The resulting matrix of chromatin interaction potential was loaded to Seurat for signature feature analysis (wilcox significance test, with adjusted p-value < 0.05) for different cell groups, and the top 5 interactions for each cell group were visualized in a heatmap (**Figure 4A**). Within the cell-type-specific chromatin interactions, one specific chromatin interaction (chr19:44,900,000-44,910,000 and chr19:44,950,000-44,960,000) connects *APOE* gene locus to 50 kb downstream. Interestingly, AD risk SNP **rs117316645** ($p < 4.8\times10^{-24}$) resides in an ATAC peak region of chr19:44,950,000-44,960,000 bin (IGV traces shown in **Figure 4B**), and is the most significant variant within this bin. Considering that *APOE* is the major genetic driver for amyloid pathology of AD, the predicted chromatin loop connecting **rs117316645** with *APOE* in astrocytes (**Figure 4C**) points to possible disrupted astrocyte function in amyloid-β clearance.

DISCUSSION

As interest in scATAC-seq continues to grow [29], there is an increasing need for efficient and accurate pre-processing and analysis software that can facilitate reproducible workflows. Our approach to scATAC-seq analysis draws on previously published tools that have been optimized for efficiency and accuracy, and should be useful for researchers grappling with increasingly large datasets. We have shown, via analysis of published 10X human PBMC and mouse brain scATAC-seq data, that scATAK compares favorably in terms of processing speed and memory usage to Cell Ranger. Furthermore, we have developed an R notebook for scATAC-seq PBMC cell clustering and cell type annotation that should be generally useful, and we have demonstrated the possibility of combined analysis of genome-wide bulk HiC type interaction map data with scATAC-seq data to calculate single-cell chromatin interaction potential matrices. Using hippocampal scATAC-seq and bulk HiChIP data from a healthy adult human brain, we presented chromatin accessibility and interaction landscapes for major brain cell types and proposed that a non-coding risk variant of Alzheimer's disease (AD) may disrupt chromatin interaction between a distal enhancer and APOE gene in astrocytes. We note that as tools are

further optimized, or when new tools are developed, it should be straightforward to replace components of our modular workflow with better alternatives.

METHODS

Software

The following software tools were used in running the scATAK workflow to generate the results and figures of this paper: kallisto (v0.46.1); bustools (v.0.40.0); minimap2 (v2.15); sambamba (v.0.7.1); Genrich; bedtools (v.2.25.0); bedGraphToBigwig. Notebooks reproducing the results and figures are available at: https://github.com/pachterlab/GP_2021_4

Hardware

All computational work was performed on a Supermicro server computer (2xXeon® Gold 6152 22-Core 2.1, 3.7GHz Turbo, 12 × 64GB Quad-Rank DDR4 2666MHz memory, 16 × 12TB Ultrastar He12 HUH721212ALE600, 7200 RPM, SATA 6Gb/s HDD) with CentOS7 operating system installed.

ACKNOWLEDGEMENTS

We thank Xun Wang for helpful suggestions. The work was possible thanks to support by the Beckman Institute at Caltech for the Caltech Bioinformatics Resource Center. FG and LP were supported in part by NIH R01 DK126925-01.

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 2019;37:925–36.
[2] Bray NL, Pimentel H, Melsted P, Pachter L. Erratum: Near-optimal probabilistic RNA-seq

quantification. Nat Biotechnol 2016;34:888.
[3] Melsted P, Booeshaghi AS, Liu L, Gao F, Lu L, Min KHJ, et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nat Biotechnol 2021;39:813–8.
[4] kb_python. Github; n.d.
[5] Giansanti V, Tang M, Cittaro D. Fast analysis of scATAC-seq data using a predefined set of genomic regions. F1000Res 2020;9:199.
[6] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094–100.
[7] Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015;31:2032–4.
[8] Gaspar JM. Genrich. Github; n.d.
[9] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.
[10] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 2010;38:576–89.
[11] Wang C, Sun D, Huang X, Wan C, Li Z, Han Y, et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome Biol 2020;21:198.
[12] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20.
[13] Fang R, Preissl S, Hou X, Lucero J, Wang X. Fast and accurate clustering of single cell epigenomes reveals cis-regulatory elements in rare cell types. BioRxiv 2019.
[14] Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods 2017;14:975–8.
[15] Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, participants in the 1st Human Cell Atlas Jamboree, et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol 2019;20:63.
[16] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326:289–93.
[17] Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods 2016;13:919–22.
[18] atac_v1_pbmc_5k -Datasets -Single Cell ATAC -Official 10x Genomics Support n.d. https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_pbmc_5k (accessed January 25, 2021).
[19] atac_v1_adult_brain_fresh_5k -Datasets -Single Cell ATAC -Official 10x Genomics Support n.d. https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k (accessed January 25, 2021).
[20] Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 2018;34:3600.
[21] Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 2015;348:910–4.
[22] Engel I, Murre C. The function of E- and Id proteins in lymphocyte development. Nat Rev Immunol 2001;1:193–9.

[23] Corces MR, Shcherbina A, Kundu S, Gloudemans MJ, Frésard L, Granja JM, et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. Nat Genet 2020;52:1158–68.
[24] Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature 2019;570:332–7.
[25] Zhou Y, Song WM, Andhey PS, Swain A, Levy T, Miller KR, et al. Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. Nat Med 2020;26:131–42.
[26] Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet 2019;51:404–13.
[27] Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 2015;16:259.
[28] Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015;523:486–90.
[29] Baek S, Lee I. Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. Comput Struct Biotechnol J 2020;18:1429–39.

Figure and Tables for

Efficient pre-processing of Single-cell ATAC-seq data

Fan Gao[1,2], Lior Pachter[1,2,3*]

[1]Division of Biology and Biological Engineering; [2]Caltech Bioinformatics Resource Center;
[3]Department of Computing & Mathematical Sciences

California Institute of Technology, Pasadena, CA, USA

*To whom correspondence should be addressed.

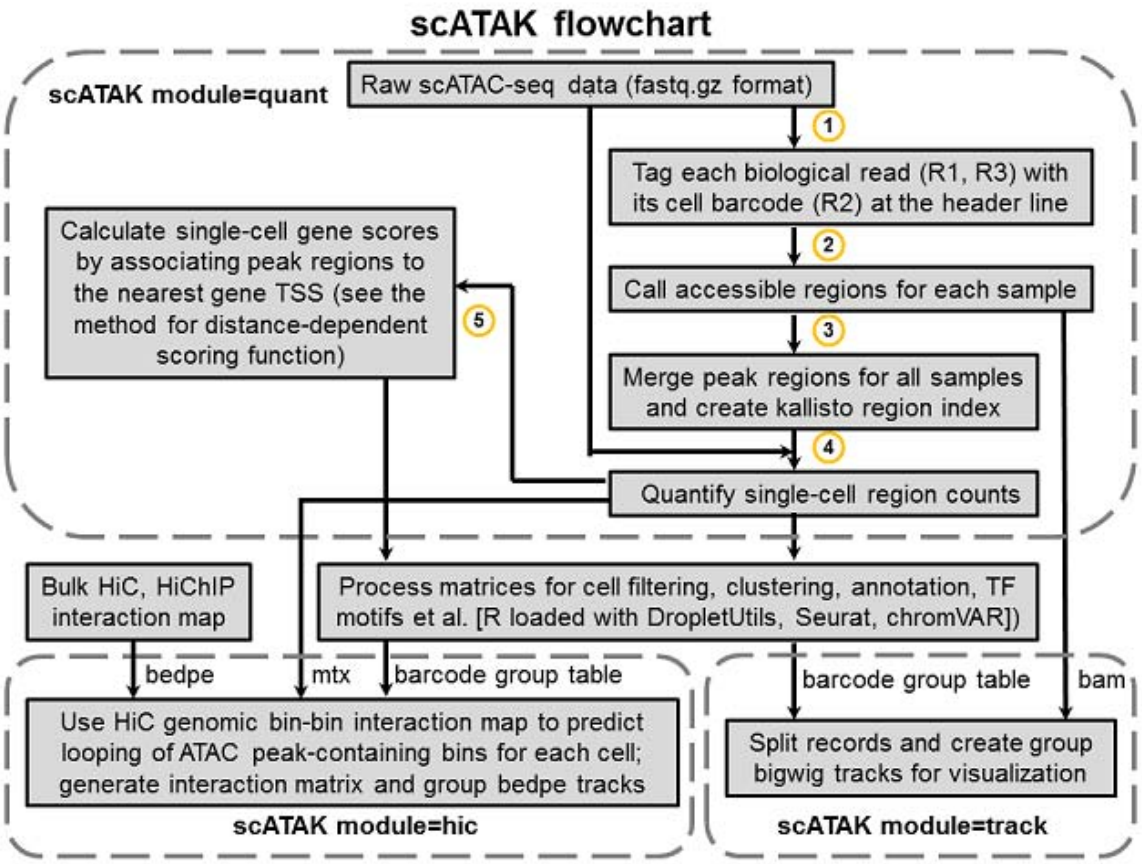

Figure 1. The scATAK workflow.

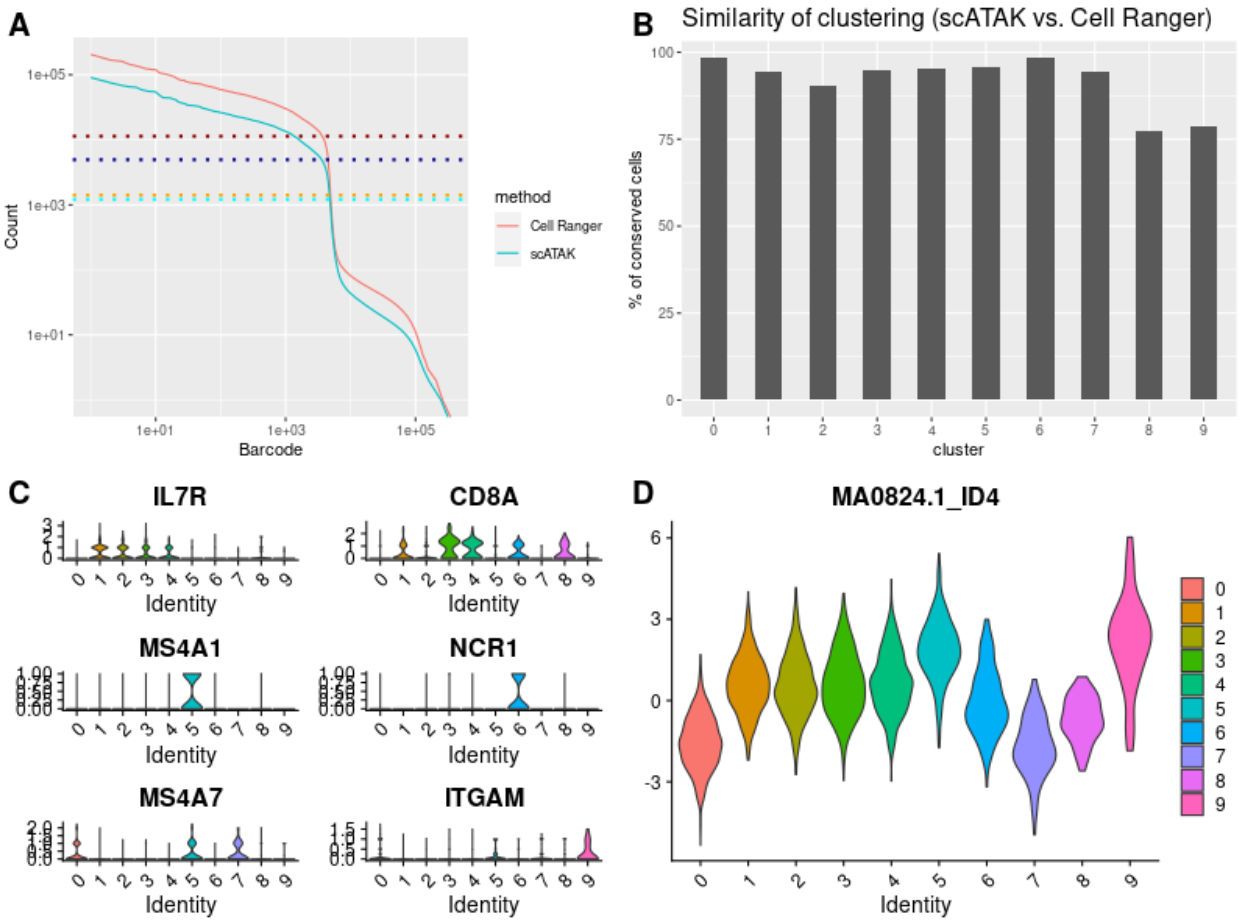

Figure 2. Benchmarking of scATAK using the human PBMC 5k data.

(A) Barcode rank plot (knee plot) showing the total UMI count vs. the rank of the barcode; (B) Percentage of overlapping barcodes between scATAK and Cell Ranger for each cell type; (C) Violin plots showing the distribution of the scATAK gene scores for known PBMC cell marker genes across different cell clusters; (D) The ID4 motif scores for the accessible chromatin regions of individual cells across cell clusters.

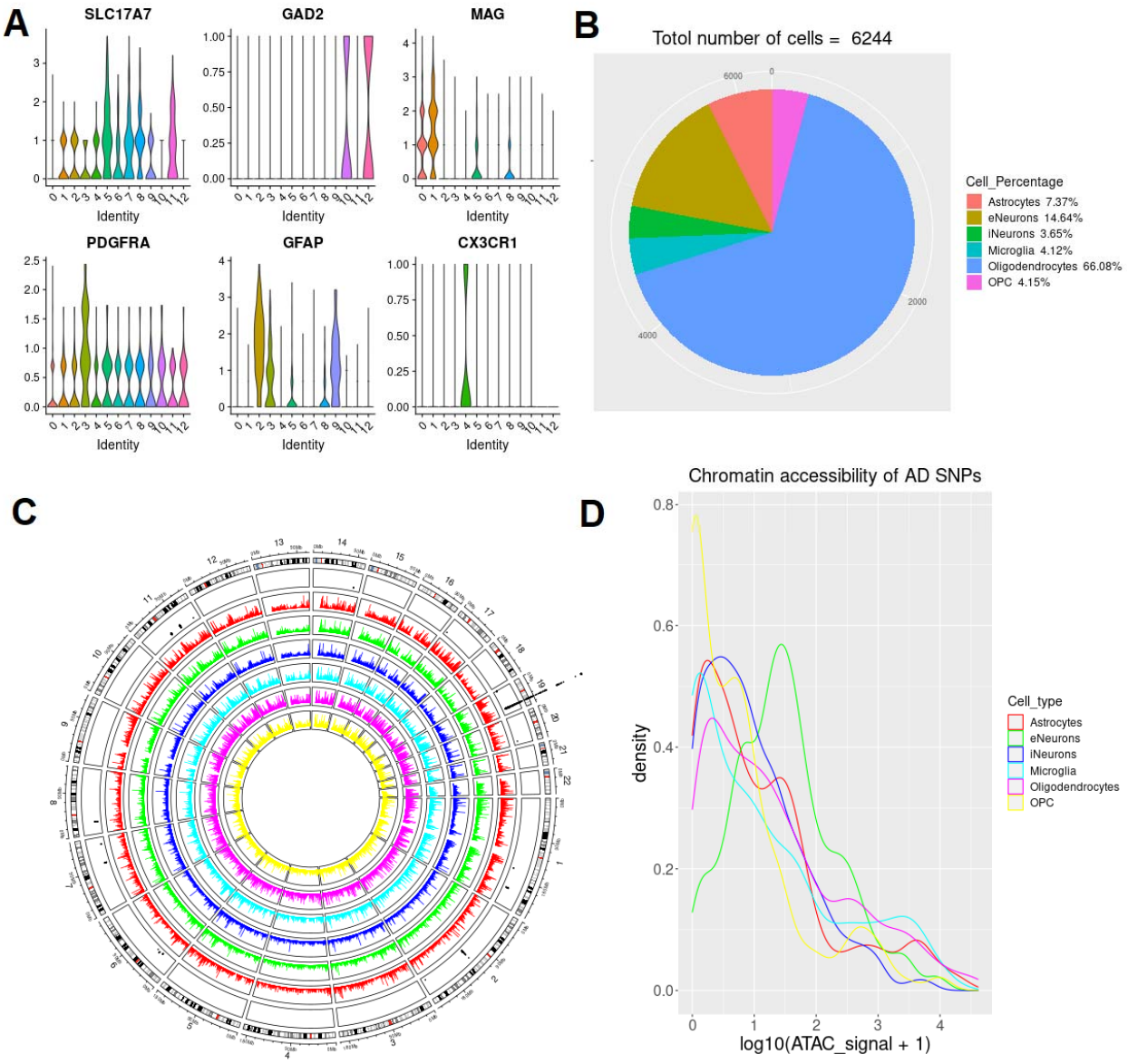

Figure 3. Cell-type-specific open chromatin landscape of the human hippocampus.

(A) Violin plots showing the distribution of the scATAK gene scores for known brain cell marker genes across different cell clusters; (B) A pie chart depicting proportions of the annotated brain cell types; (C) A circos plot visualizing AD GWAS loci together with open chromatin landscapes for different cell types; (D) A density plot for the calculated cell-type-specific open chromatin signals of AD GWAS SNPs).

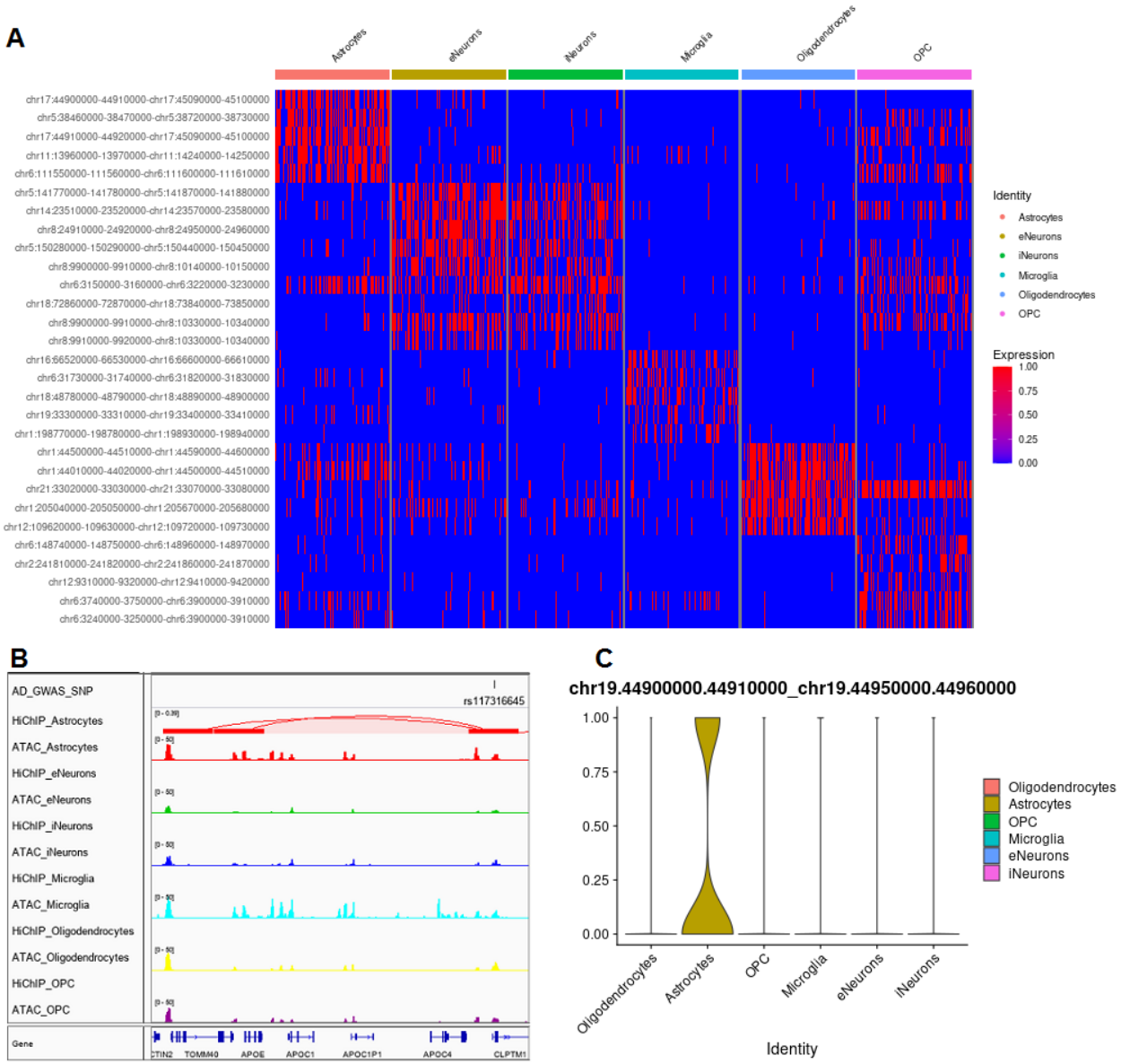

Figure 4. Predicted cell-type-specific chromatin interactions and connection between the AD risk variant **rs117316645** and the *APOE* gene.

(A) The top 5 predicted signature chromatin-chromatin interactions for each cell type; (B) IGV browser view of predicted cell-type-specific chromatin loop and open chromatin tracks around *APOE* gene locus; (C) Cell type specificity of a chromatin looping that links **rs117316645** to the *APOE* gene in astrocytes.

PBMC Total read pairs: 224,636,372

CPU threads	Real time (scATAK)	Real time (Cell Ranger)	Fold change
8	64 min	1171 min	18.3
4	81 min	1955 min	24.1
2	139 min	3456 min	24.9

CPU threads	Max RSS (scATAK)	Max RSS (Cell Ranger)	Fold change
8	8.66 GB	12.99 GB	0.67
4	8.66 GB	9.31 GB	0.93
2	8.66 GB	8.35 GB	1.04

Adult mouse brain Total read pairs: 244,056,346

CPU threads	Real time (scATAK)	Real time (Cell Ranger)	Fold change
8	72 min	1045 min	14.5
4	98 min	1752 min	17.9
2	164 min	3027 min	18.5

CPU threads	Max RSS (scATAK)	Max RSS (Cell Ranger)	Fold change
8	7.86 GB	12.17 GB	0.65
4	8.46 GB	9.79 GB	0.86
2	8.03 GB	9.58 GB	0.84

Running time breakdown in seconds (scATAK, CPU threads = 8)

Sample ID	PBMC	Adult mouse brain
genome indexing	61	55
genome alignment and peak calling	2823	3243
kallisto indexing	47	127
kallisto / bustools quantification	892	958
gene activity quantification	86	74

Table 1. Comparison of running time for the scATAK and CellRanger pipelines.

Sample ID	PBMC	adult mouse brain
Processed reads	224,636,372	244,056,346
Pseudoaligned reads	100,549,039	118,667,309
Pseudoalignment rate %	44.76%	48.62%
Pseudoaligned reads in the whitelist	90,210,039	114,063,492
Whitelist read rate %	89.72%	96.12%
Pseudoaligned reads with BC corrected	4129483	1,504,416
Correction rate %	4.11%	1.27%

Table 2. Read pseudoalignment statistics for kallisto *bus* and bustools.