World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:15, No:10, 2021

Efficient Pre-Processing of Single-Cell Assay for Transposase Accessible Chromatin with High-Throughput Sequencing Data

Fan Gao, Lior Pachter

***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 15 to 18 times faster than Cell Ranger on mouse and human samples. Our tool can also calculate chromatin interaction potential matrices and generate open chromatin signal and interaction traces for cell groups. We use scATAK tool to explore the chromatin regulatory landscape of a healthy adult human brain and unveil cell-type specific features, and show that it provides a convenient and computational efficient approach for pre-processing single-cell ATAC-seq data.

***Keywords*—**Single-cell, ATAC-seq, bioinformatics, open chromatin landscape, chromatin interactome.

I. 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 consist 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].

More recently, we incorporated kallisto, bustools and other tools into an scATAC-seq software suite, namely scATAK, to process scATAC-seq data. In this article, we provided a schematic overview of this pipeline; we compared processing speed and memory usage of scATAK with Cell Ranger using published 10X human PBMC and mouse brain scATAC-seq data; we further created an R notebook for PBMC cell clustering and cell type annotation; finally, we presented a unique feature of scATAK that combines genome-wide bulk HiC type interaction map with scATAC-seq matrix to calculate single-cell chromatin interaction potential matrix. 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.

II. Results

A. Overview of scATAK and Benchmarking

An overview of scATAK procedure is illustrated in Fig. 1. As noted, scATAK is a command-line driven 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 [5], converted to a name sorted BAM alignment file using Sambamba [6], and then subject to peak calling using Genrich [7]; 3) Called peak regions for all samples are merged to generate a list of accessible chromatin regions using bedtools [8] for creating a kallisto index file; 4) With kallisto accessible region index as reference, raw scATAC-seq files are revisited to generate 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 d from ATAC-seq peak centers to transcription start sites (*TSS*) and associate peak regions to the nearest gene *TSS*, a strategy similar to HOMER [9] peak annotation. Activity score S for gene i is calculated as weighted sum of associated peaks P, with $Si = \Sigma\, Wij \times Pj$, where W is a distance-dependent step function for weight, values from 1 ($d \leq$ 2 kb), 0.7 (2 kb < $d \leq$ 5 kb), 0.5 (5 kb < $d \leq$ 10 kb), 0.25 (10 kb < $d \leq$ 20 kb) to 0.03 (20 kb < $d \leq$ 50 kb). A distance dependent weight is originally proposed in MAESTRO pipeline [10] to better model gene activity. Instead of using computationally

Fan Gao is with Division of Biology and Biological Engineering and with Caltech Bioinformatics Resource Center.

Lior Pachter is with Division of Biology and Biological Engineering, Caltech Bioinformatics Resource Center and with Department of Computing & Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA (corresponding author, e-mail: lpachter@caltech.edu).

expensive exponential decay to calculate *W*, we simply employed a step function to speed up processing. With accessible region and gene activity count matrices, further analyses can be performed within R or Python notebooks using secondary analyses tools like Seurat [11], snapATAC [12] or chromVAR [13]. For our proof-of-principle analyses, we created R notebooks with DropletUtils [14] and Seurat loaded. After cell clustering and annotation are completed, scATAK *track* module uses cell barcode — cell group table together with pseudo-bulk ATAC-seq alignment file 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 known bulk HiC [15] or HiChIP [16] interactome map together with 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.

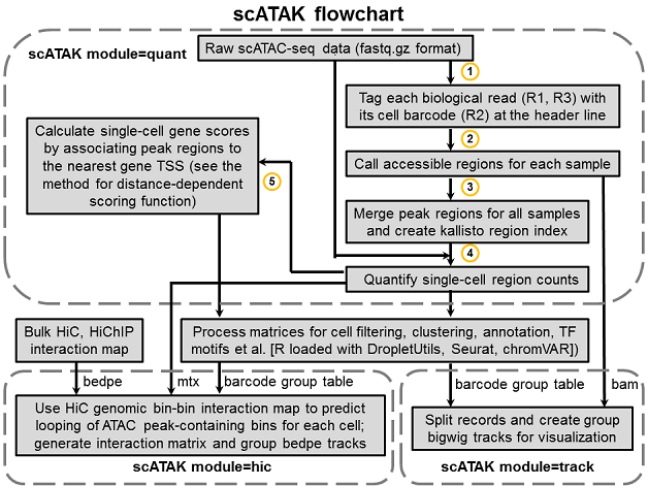

Fig. 1 Schematic flowchart of scATAK pipeline

For benchmarking purposes, we downloaded two 10X scATAC-seq datasets that are hosted on 10X Genomics data links [17], [18]. In brief, 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 CellRanger (atac-1.2.0) software. With 2, 4, 8 CPU threads, real run time and memory usage were monitored by snakemake pipeline [19]. As shown in Table I, when PBMC data were processed, scATAK *quant* was roughly 17, 23, 25 times faster than Cellranger with 8, 4, 2 CPU threads employed. For mouse brain data, scATAK was about 14, 16, 18 times faster when using 8, 4, 2 CPU threads. 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 II, kallisto *bus* pseudo-alignment method works well for mapping raw reads to ATAC-seq peak regions, with 45% and 49% pseudo-alignment rate 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, Fig. 2 A). As noted, 3,528 cell barcodes were shared between 3,595 filtered barcodes from scATAK and 3,653 filtered barcodes from Cell Ranger, suggesting 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 same default settings of Seurat (see method) both resulted in 12 cell clusters, visualized in UMAPs (Figs. 2 C, D). Cell barcodes within each scATAK generated cluster were subject to overlap statistical analysis (Fisher's exact test) with clustered barcodes from Cell Ranger, with -log10P-value visualized in a heatmap (Fig. 2 B). Clearly, cell clusters from two different pipelines are highly concordant with each other.

TABLE I
COMPARISON OF REAL RUNNING TIME FOR SCATAK AND CELLRANGER PIPELINES

Sample ID	Total read pairs		
PBMC	224,636,372		
CPU threads	Real time (scATAK)	Real time (CellRanger)	Fold increase
8	64 min.	1171 min.	18.3
4	81 min.	1955 min.	24.1
2	139 min.	3456 min.	24.9
Sample ID	Total read pairs		
Adult mouse brain	244,056,346		
CPU threads	Real time (scATAK)	Real time (CellRanger)	Fold change
8	72 min.	1045 min.	14.5
4	98 min.	1752 min.	17.9
2	164 min.	3027 min.	18.5

TABLE II
STATISTICS FROM KALLISTO *BUS* AND BUSTOOLS

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%

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 (Fig. 2 E). 12 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 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, the top signature motif for B cells is MA0824.1_ID4 (Fig. 2 F). This observation is consistent with regulatory roles of Id proteins in lymphocyte development [20]. Overall, secondary analyses using pre-processed results from the scATAK pipeline revealed expected biological insight from PBMC cells.

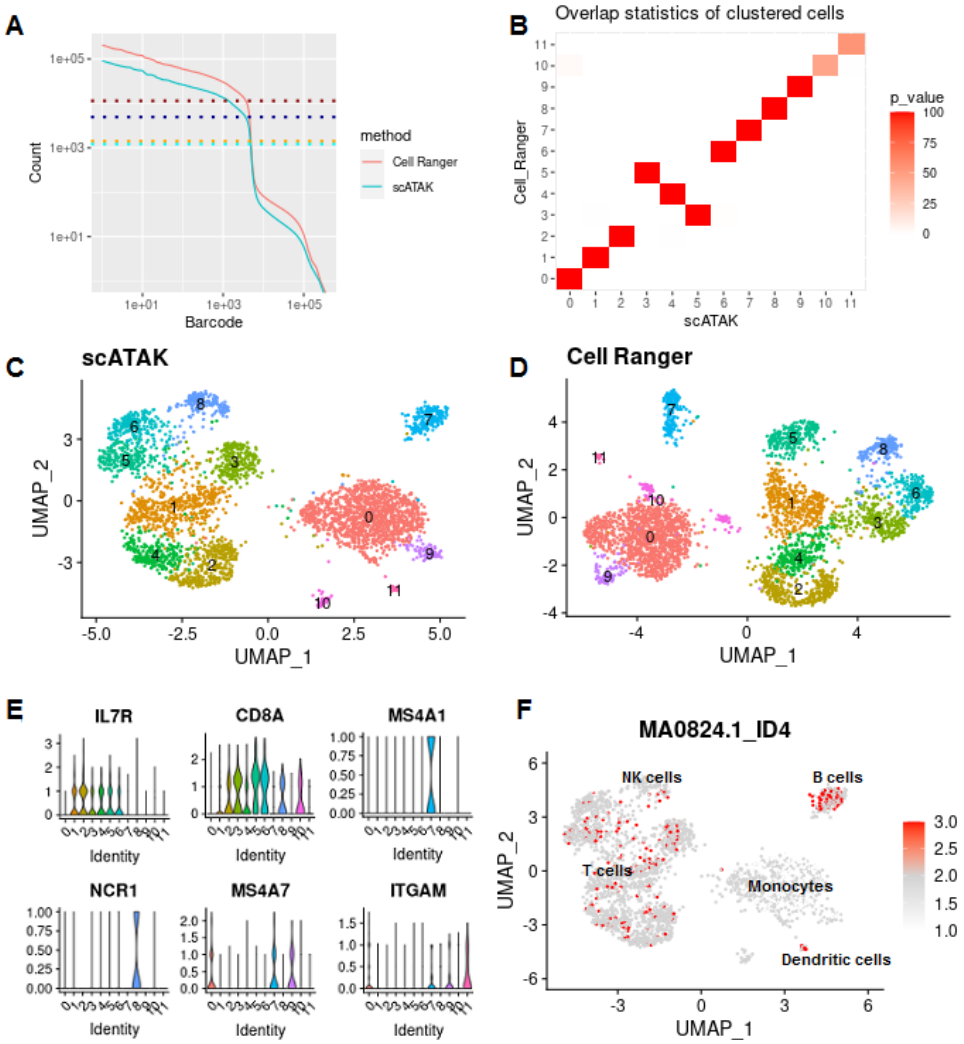

Fig. 2 Benchmarking using PBMC data

B. Exploration of Chromatin Accessibility and Interaction Landscapes in Human Brain

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 could be postulated by high-resolution genome-wide chromatin interactome map, we still lack cell-type specificity of the interactions. As mentioned above, scATAK implemented 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 showed HiChIP [16] – 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 exploration, 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 [21, GEO accession numbers GSM4441823 and GSM4441836]. Total of 6,082 cell nuclei were recovered from DropletUtils and 13 cell clusters were generated using Seurat and visualized in a UMAP (Fig. 3 A). 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) (Fig. 3 B), six major brain cell types were assigned (Fig. 3 A). Noted from the UMAP, excitatory neurons have at least two separated sub-clusters, with clusters 6, 7, 8, 11 forming one sub-group and cluster 5 forming the other sub-group. This complex structure of open chromatin landscape in excitatory neurons is consistent with multi subtypes of excitatory neurons observed from brain single-cell RNAseq data [22], [23], demonstrating chromatin accessibility as another molecular marker for sub-clustering of excitatory neurons.

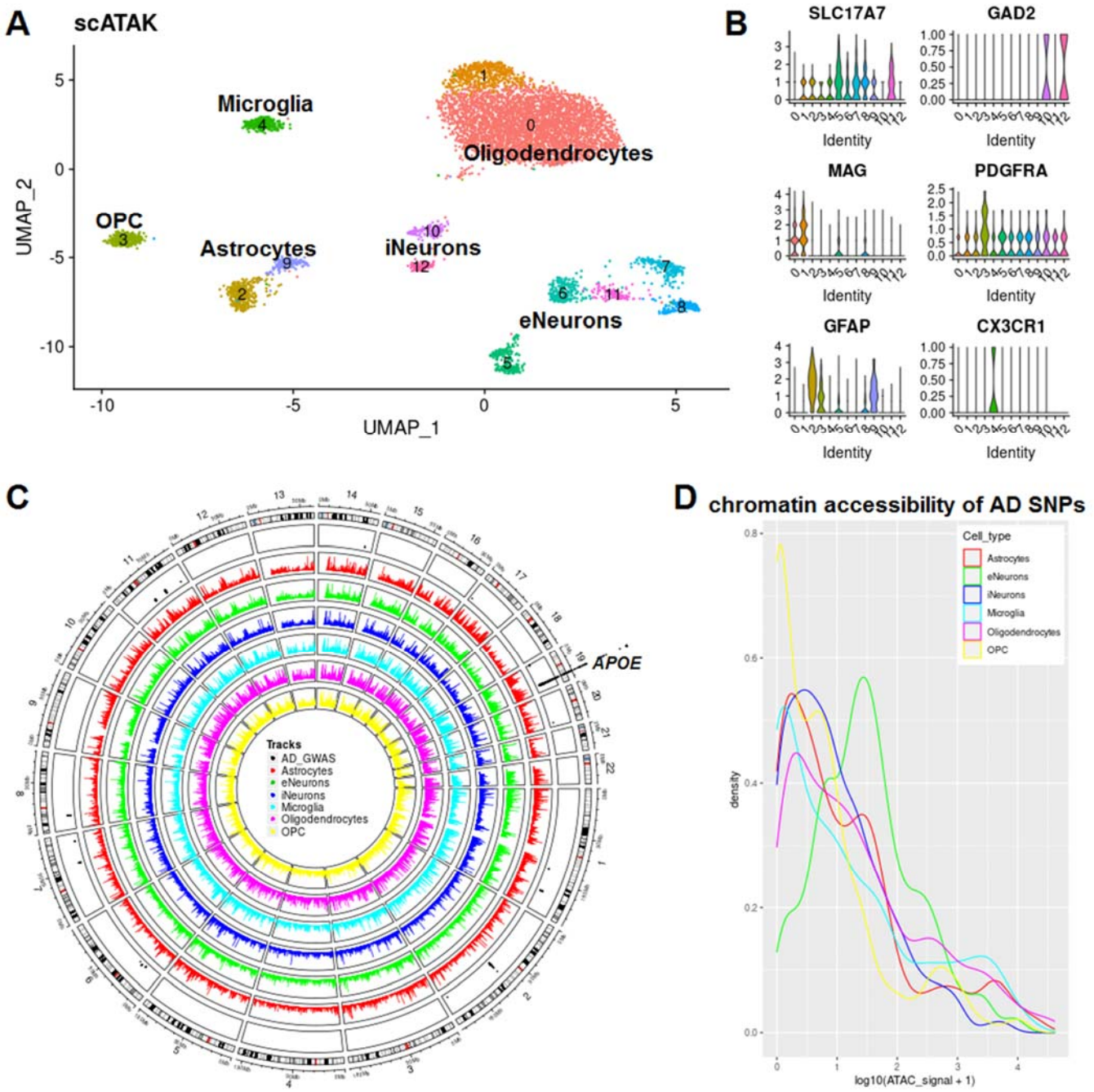

Fig. 3 Cell type specific open chromatin landscape of human hippocampus

We next asked how genetic variants could explain susceptibility of hippocampal cells to AD. The scATAK *track* module generated group ATAC signal tracks (normalized by mapped group read counts) from cell barcode – cell group table and sample pseudo-bulk alignment file. A circos plot (Fig. 3 C) provided a gnome-wide view of human GWAS AD risk SNPs [24], [25] (SNPs with $p < 1\times10^{-9}$ included) and ATAC signals in different cell types (signals binned for every 200 kb genomic window). AD risk SNPs were further associated with 2 kb genomic bins to calculate chromatin accessibility in different cell types. Shown in Fig. 3 D density plot, astrocytes, microglia and oligodendrocytes are enriched with subsets of SNP regions

that are highly accessible ($\log_{10}$(ATAC-signal + 1) > 3). This observation suggests that these cell types are vulnerable to AD associated genetic variation.

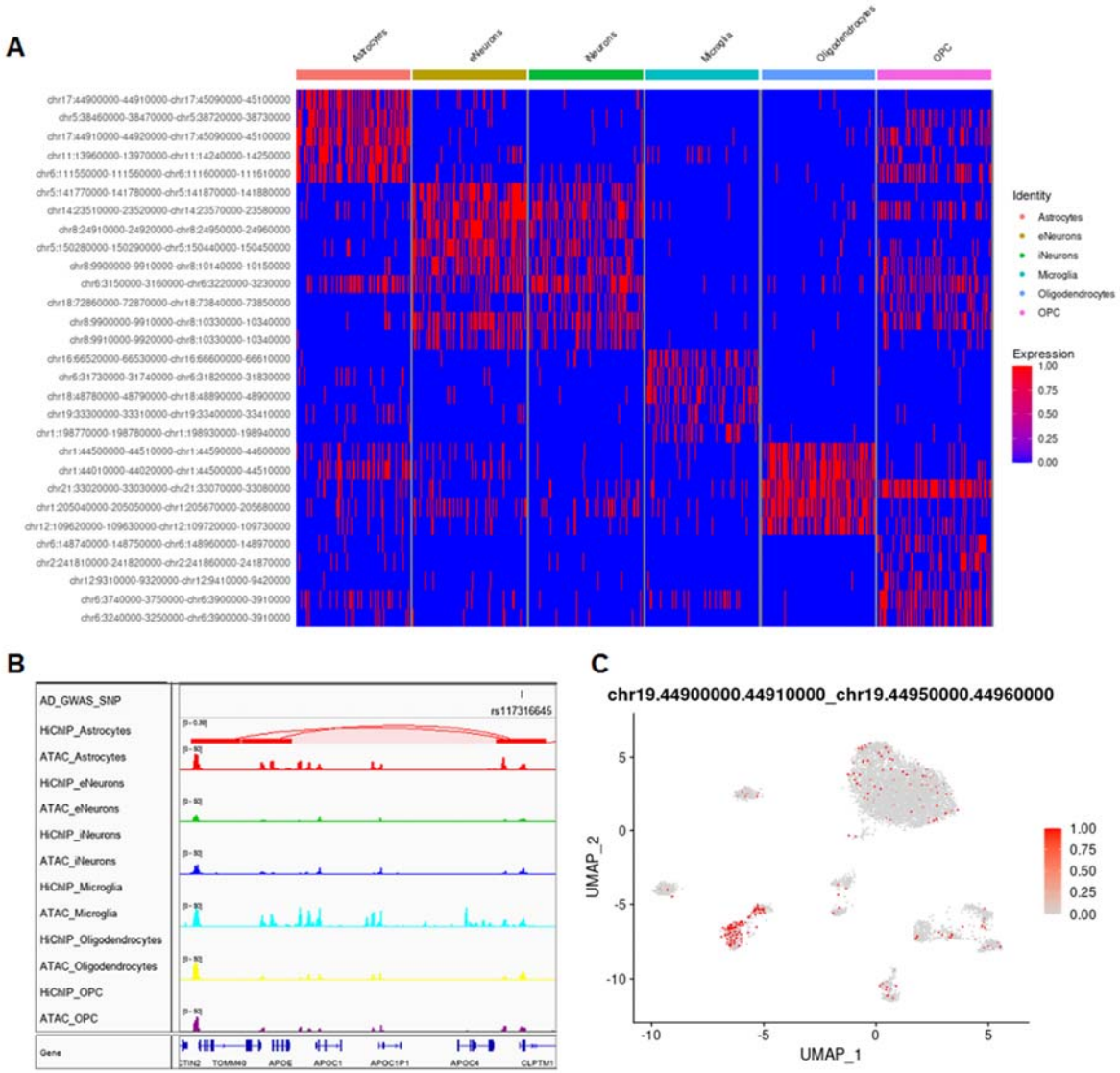

Fig. 4 Predicted cell-type specific chromatin interactions and connection between AD risk variant rs117316645 and *APOE* gene

We next loaded the scATAK *hic* module to subset genomic looping bin pairs (10 kb resolution interaction map) identified from bulk histone H3K27ac HiChIP data (GSM4441836) using single-cell chromatin accessibility map already created in scATAK *quant* step. As note, the downloaded map was generated by HiC-Pro [25] to include cis-interactions between 20 kb and 2 Mb. 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 [26]. For *N* accessible regions in a single cell, $N\times(N\text{-}1)/2$ 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 top 5 interactions for each cell group were visualized in a heatmap (Fig. 4 A). 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 Fig. 4 B), and is the most significant variant within this bin. Considering *APOE* is the major genetic driver for amyloid pathology of AD, the predicted chromatin loop connecting

rs117316645 with *APOE* in astrocytes (Fig. 4 C) postulates disrupted astrocyte function in amyloid-β clearance that could be further tested.

In summary, we demonstrated the feasibility of using kallisto/bustools based scATAK *quant* module to quickly pre-process scATAC-seq data. Compared to industry standard Cell Ranger, our pipeline is up to 25 times faster, enabling researchers with restricted computational resources to tackle the raw data. An additional scATAK *track* module bridges between matrix analysis and genome-wide data visualization. The unique scATAK *hic* module attempts to connect single-cell chromatin accessibility to active 3D genome interactome. With modular design, scATAK pipeline can easily communicate with other scATAC analysis tools. Finally, the source codes and benchmarks of scATAK are freely available to the scientific community.

III. Discussions

Kallisto pseudoalignment method was originally developed to rapidly process RNA-seq data and quantify transcript abundance. Together with bustools, the workflow is extremely efficient to pre-process single-cell RNA-seq data. To pre-process scATAC data, the initial design was to index the entire genome for kallisto pre-processing. It turned out that this workflow is computationally too expensive. Then a smaller index file was created for curated gene regulatory regions only (promoters, gene body, enhancers etc.) to speed up kallisto pre-processing. However, this approach could not capture the most important feature – accessible "peak" regions from ATAC-seq data. Thus, minimap2 and Genrich, two efficient tools for genome mapping and ATAC-seq peak calling, were included in our scATAC-seq workflow as pre-kallisto steps. With 45% (PBMC) and 49% (mouse brain) raw reads pseudo-aligned to the "peak" regions, this workflow efficiently extracts biologically important features from raw scATAC data.

IV. Methods

A. Software

The following software tools were included in scATAK pipeline: kallisto (v0.46.1); bustools (v.0.40.0); minimap2 (v2.15); sambamba (v.0.7.1); Genrich; bedtools (v.2.25.0); bedGraphToBigwig.

B. Hardware

All computational work was performed on a Supermicro server computer with CentOS7 operating system installed.

V. Author Contributions

FG and LP designed the project. FG wrote the pipeline and performed data analysis. FG and LP wrote the manuscript.

References

[1] Satpathy A. T., Granja J. M., Yost K. E., Qi Y., Meschi F., McDermott G. P., 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 N. L., Pimentel H., Melsted P., Pachter L. Erratum: Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 2016;34:888.
[3] Melsted P., Sina Booeshaghi A., Gao F., Beltrame E., Lu L., Hjorleifsson KE, et al. Modular and efficient pre-processing of single-cell RNA-seq. Cold Spring Harbor Laboratory 2019:673285. https://doi.org/10.1101/673285.
[4] kb_python. Github; https://github.com/pachterlab/kb_python
[5] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094–100.
[6] Tarasov A., Vilella A. J., Cuppen E., Nijman I. J., Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015;31:2032–4.
[7] Gaspar J. M. Genrich. Github; https://github.com/jsh58/Genrich
[8] Quinlan A. R., Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] Schep A. N., Wu B., Buenrostro J. D., Greenleaf W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods 2017;14:975–8.
[14] Lun A. T. L., Riesenfeld S., Andrews T., Dao T. P., 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.
[15] Lieberman-Aiden E., van Berkum N. L., 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.
[16] Mumbach M. R., Rubin A. J., Flynn R. A., Dai C., Khavari P. A., Greenleaf W. J., et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods 2016;13:919–22.
[17] 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).
[18] 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).
[19] Köster J., Rahmann S.. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 2018;34:3600.
[20] Engel I., Murre C. The function of E- and Id proteins in lymphocyte development. Nat Rev Immunol 2001;1:193–9.
[21] Corces M. R., Shcherbina A., Kundu S., Gloudemans M. J., Frésard L., Granja J. M., 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.
[22] Mathys H., Davila-Velderrain J., Peng Z., Gao F., Mohammadi S., Young J. Z., et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature 2019;570:332–7.
[23] Zhou Y., Song W. M., Andhey P. S., Swain A., Levy T., Miller K. R., 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.
[24] Jansen I. E., Savage J. E., Watanabe K., Bryois J., Williams D. M., 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.
[25] Servant N., Varoquaux N., Lajoie B. R., 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.
[26] Buenrostro J. D., Wu B., Litzenburger U. M., Ruff D., Gonzales M. L., Snyder M. P., et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015;523:486–90.