Integrated analysis of chromatin accessibility and gene expression at the single-nucleus level in a cancer-immune cell mixture

Jieun Seo

The Graduate School
Yonsei University
Department of Medical Science

Integrated analysis of chromatin accessibility and gene expression at the single-nucleus level in a cancer-immune cell mixture

A Master's Thesis Submitted
to the Department of Medical Science
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Master's of Medical Science

Jieun Seo

January 2025

This certifies that the Master's Thesis of Jieun Seo is approved

Thesis Supervisor : Hyoung-Pyo Kim

Thesis Committee Member : Hyun Seok Kim

Thesis Committee Member : Byungjin Hwang

The Graduate School
Yonsei University

January 2025

ACKNOWLEDGEMENTS

I would like to sincerely thank everyone who supported me throughout the course of this research. First and foremost, I would like to thank Professor Hyoung-pyo Kim for guiding me to have a new and broad perspective throughout the entire research process and for granting me with the opportunity to conduct research in the field of epigenetics. Based on the professor's insightful guidance and suggestions regarding the expansiveness of research, I was able to identify various new possibilities from different perspectives and apply them to this study. I would also like to sincere gratitude to Professors Hyun Seok Kim and Byungjin Hwang for their practical and realistic advice, which greatly facilitated the progress of this study.

In addition, I am deeply grateful to my fellow laboratory members who worked alongside me and provided tremendous support to help me adapt to a new field of study. Special thanks go to MiKyung Kim, who helped me establish a solid foundation in the unfamiliar field of bioinformatics and kindly answered even the smallest questions; to Chul Min Yang, who guided me to reflect constantly on the purpose and significance of the research results; to Eun-Chong Lee, who provided sharp insights and meticulous guidance to ensure no details were overlooked, and taught me how to think critically for myself; to Jung-Sik Joo, who provided new perspectives and logical clarity on the uncertainties of the research and offered invaluable advice and encouragement, giving me the strength to persevere; to Kyungwoo Kim, who offered clear and fundamental explanations on methodological issues in analysis and helped me build a solid foundation as an analyst; and to Heon-Woo Kwon and Chang Hoon Yang, who not only shared valuable opinions and ideas while conducting this research together, but also entered the lab at the same time as me, offering mutual encouragement and support. I am also deeply thankful to Hyunseong Kim, who laid the methodological groundwork for this research and generously shared his knowledge; to Hyejin Kim, who created a warm and encouraging environment in the lab with her kind support; to Sugyung Kim, who provided practical assistance and shared her insights with me; and to Chanyeon Park, who encouraged me to try new analytical approaches and explore a variety of possibilities.

Lastly, I am profoundly grateful to my beloved family and friends who have always supported my path unconditionally, offering endless encouragement and love. Their unwavering support has been a constant source of motivation throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Integrated analysis of chromatin accessibility and gene expression at the single-nucleus level in a cancer-immune cell mixture

Understanding cellular heterogeneity is crucial for unraveling the complexity of tissue function and disease progression. SHARE-seq, a single-cell multiomics technology, provides an opportunity to explore the epigenomic and transcriptomic landscapes at the level of individual cells, surpassing previous approaches that average profiles across populations of cells. In this study, a bioinformatics analysis pipeline for investigating epigenomic heterogeneity was validated using public SHARE-seq data and applied to an in-house SHARE-seq dataset from a mixture of cancer and immune cells to identify cellular heterogeneity. To verify the reproducibility of the pipeline, publicly available SHARE-seq data from human kidney tissue were used. This analysis successfully reconstructed the transcriptomic and epigenomic heterogeneity of various cell types from kidney, identifying clear cell clusters based on transcriptomic and chromatin accessibility profiles that aligned with results from previous studies. Furthermore, the validated pipeline was utilized to integratively analyze the two modalities of an in-house SHARE-seq dataset from a mixture of a colorectal cancer cell line and an immune cell line, successfully distinguishing the two cell types. Additionally, we identified regulatory chromatin regions with strong correlations between the two modalities and analyzed their associations with super-enhancer regions. This revealed that chromatin accessibility and gene expression are differentially regulated depending on the cell type, and factors such as the degree of peak-gene association and accessibility levels can collectively influence the expression of cell type-specific genes. Moreover, the analysis highlighted that the activity of transcription factors varies across cell types, affecting the expression of genes that have cell type-specific functions. This reaffirms that the complex interplay among chromatin accessibility, gene regulation, and transcription factor activity collectively contributes to defining cell type-specific identities.

In conclusion, this study comprehensively explored the relationship between gene expression and chromatin accessibility through the integrative analysis of single-cell multiomics data, identifying key regulatory elements that define cellular identity and function. These findings are expected to provide critical insights into the regulatory environment and heterogeneity of individual cells in diseases such as cancer, advancing our understanding of cellular mechanisms at a single-cell resolution.

Key words : cellular heterogeneity, single-cell multiomics, SHARE-seq, bioinformatics, integrated analysis

1. Introduction

Tissues in multicellular organisms perform diverse functions, and the cells that make up these tissues share almost identical genome sequences but exhibit distinct gene expression patterns, enabling them to carry out different cellular functions. The regulation of gene expression and the resulting cellular heterogeneity are actively studied, particularly due to their important roles in diseases such as cancer[1-4]. Gene expression is initiated through the transcription of genomic DNA into messenger RNA (mRNA), a process that can be controlled by the interaction of proteins, including transcription factors and initiators, with cis-regulatory elements such as promoters and enhancers[5-7]. Additionally, cellular heterogeneity in gene expression arises from epigenetic features such as nucleosome positioning and composition, histone tail modifications, and three-dimensional structural interactions[8-10]. Therefore, a comprehensive understanding of gene expression heterogeneity requires investigating the interplay of these various regulatory mechanisms.

Bulk-cell experiments, which analyze large populations of cells simultaneously, provide an aggregate signal representing the cell population. These methods are insufficient for distinguishing cellular differences in transcriptomic and epigenetic features. However, advances in single-cell multi-omics research have overcome this limitation by enabling the analysis of transcriptomes for each cell within a sample[11]. Single-cell sequencing is particularly valuable for identifying rare cell types that are difficult to identify in bulk sequencing, thereby helping to optimize therapeutic strategies for issues such as tumor formation and therapy resistance[12].

Nevertheless, unimodal single cell technologies can only reveal cellular heterogeneity for individual epigenetic and transcriptomic features and cannot simultaneously profile multiple ones, chromatin accessibility, and gene expression within the same single cell. As a result, while these methods suggest potential correlations between epigenetic phenomena and transcription levels, they cannot directly investigate these relationships[13]. To address this limitation, multimodal single-cell sequencing technologies have been introduced that can simultaneously analyze gene expression and additional aspects of chromatin state.

Simultaneous high-throughput ATAC and RNA expression sequencing (SHARE-seq)[14] is a technique that allows for the investigation of both epigenomic and transcriptomic dynamics from the same cell. This approach enables large-scale, cost-effective measurements of chromatin accessibility and gene expression in single cells, either individually or jointly. Through SHARE-seq, accurate cell type definition can be achieved by elucidating the correlations between chromatin accessibility and gene expression. By leveraging cellular heterogeneity, this method also can infer chromatin accessibility and transcription relationships and identify high-density peak-gene associations, known as domains of regulatory chromatin (DORCs).

This study focuses on integrating multimodal datasets for analyzing transcriptome and chromatin accessibility from the same cell simultaneously. Specifically, this study validates and applies bioinformatic analysis methods for SHARE-seq data. By applying it to in-house SHARE-seq data from a mixed sample of two heterogeneous cell lines, this study aims to identify the unique biological characteristics of each cell line by separating them and investigate the relationships between regulatory chromatin and gene expression that can define the distinct identity of each cell line.

2. Materials and methods

2.1. Datasets

2.1.1. SHARE-seq data from a mixed sample of HCT116 and NK92 cell lines

In-house SHARE-seq data from a mixed sample of HCT116 and NK92 cell lines was generated by Heon-Woo Kwon using protocol published at Ma et al. 2020[14] and optimized by Ph.D. Chul Min Yang and Ph.D. Eun-Chong Lee.

2.1.2. SHARE-seq data from human kidney tissue

SHARE-seq data from human kidney tissue was obtained from GSE234788[15].

2.1.3. H3K27ac CUT&Tag of HCT116 cells

H3K27ac CUT&Tag data from HCT116 cell line was generated by Heon-Woo Kwon.

2.1.4. H3K27ac ChIP-seq of NK92 cells

H3K27ac ChIP-seq data from NK92 cells was obtained from GSE227664[16].

2.1.5. Bulk RNA-seq and RUNX2 ChIP-seq data of human NK cells

Intersected gene list data of bulk RNA-seq and RUNX2 ChIP-seq from human NK cells in each condition of RUNX2 knockdown and RUNX2 overexpression were obtained from Wahlen et al. 2022[17].

2.2. Bulk sequencing data processing

2.2.1. CUT&Tag analysis

The adapter sequences from paired-end sequencing reads which are 101 bp were trimmed using trim_galore[18] (v0.6.10). Processed reads subsequently were aligned to the hg38 reference genome using bowtie2[19] (v2.5.1) with --local. Duplicate reads were marked using Picard[20] (v2.26.0) with default parameters and duplicates, mitochondrial reads, and low-quality reads were filtered out using SAMtools[21] (v1.17) with -q 30 –F 1804 –f 2. The preceding analyses were individually performed for H3K27ac CUT&Tag sample and input data from the HCT116 cell line. Since H3K27ac is a narrow histone mark, narrowpeaks were identified using callpeak in MACS2[22] (v 2.2.7.1) with -g hs -f BAMPE --nomodel -q 0.05 and input reference signal data. Peaks located in blacklist and patch regions were filtered out.

2.2.2. ChIP-seq analysis

The adapter sequences from paired-end sequencing reads which are 101 bp were trimmed using trim_galore. Processed reads subsequently were aligned to the hg38 reference genome using bwa[23] (v0.7.17) with default parameter settings. Duplicate reads were marked using Picard with default parameters and duplicates, mitochondrial reads, and low-quality reads were filtered out using SAMtools with -q 30 –F 1804 –f 2. The preceding analyses were individually performed for H3K27ac ChIP-seq sample and input data from the NK92 cell line. Narrowpeaks were identified using callpeak in MACS2 with -g hs -f BAMPE --nomodel -q 0.05 and input reference signal data. Peaks located in blacklist and patch regions were filtered out.

2.3. SHARE-seq data processing

2.3.1. SHARE-seq data pre-processing

SHARE-seq data (.fastq.gz) was pre-processed using scripts previously described and available at https://github.com/masai1116/SHARE-seq-alignmentV2/[14]. Gene annotation and sequence files from the GENCODE website[24] were used. The Genome Reference Consortium Human Build 37 patch release 13 (GRCh37.p13; hg19) was used for analyzing SHARE-seq data of human kidney tissue, while the Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13; hg38) was used for in-house SHARE-seq data of the mixed sample of HCT116 and NK92 cell lines. Barcode demultiplexing was performed allowing one mismatch based on the introduced barcodes in split-pool barcoding. Reads with disabled adapters and low-quality sequences were trimmed using fastp[25] (v0.23.4). For snRNA-seq (SHARE-RNA) data, due to the characteristic presence of polyA tails in mRNA, the read2 sequences were excluded, and only the read1 FASTQ file was aligned to the reference genome using STAR[26] (v2.5.2b). The number of reads mapped to genomic regions was quantified using FeatureCount[27] (v2.0.6), and unique UMI-based read grouping was performed using UMI-tools[28] (v1.1.5) to obtain unique reads by removing duplicated reads. The alignment of snATAC-seq (SHARE-ATAC) data was conducted using bowtie2 (v2.5.3). Reads that were unmapped, not primarily aligned, or aligned to chrM or chrY were removed. Barcodes with fewer than 100 reads in the SHARE-seq data from human kidney tissue and fewer than 50 reads for SHARE-seq data from the mixed sample of HCT116 and NK92 cell lines were filtered out. The read distribution was checked using RSeQC[29] (v5.0.2). This process produced a count matrix (.h5 file) representing gene expression and a fragment profile (.bed file) for each individual cell. The bigWig files were generated by bamCoverage in deepTools (v3.5.5) with the --normalizeUsing CPM option.

The .h5 files were processed to generate count matrices using the scanpy.read_10x_h5 function from Scanpy[30] (v1.9.8). For SHARE-seq data from human kidney tissue, count matrices from data sequenced on different NovaSeq flowcells were combined using the anndata.concat function. For the processing of the .bed files for snATAC-seq analysis, Tabix[31] (v1.20) was used to merge all fragment profiles of human kidney tissue from different flowcells, and the CreateFragmentObject function of Signac[32] (v1.14.0) was used to create a single object from the fragment profile. Peaks were identified using the CallPeaks function (extsize=150), which utilizes MACS2 (v2.2.9.1), and a count matrix was generated using the FeatureMatrix function resulting in the identification of a total of 189,184 features for human kidney data and 184,399 features for in-house mixed sample data.

2.3.2. snRNA-seq data analysis

All snRNA-seq analysis was executed on Scanpy. For SHARE-RNA data from human kidney tissue, all key parameters were followed as outlined in the paper[15]. In brief, cells with fewer than 200 or more than 5,000 genes detected, as well as cells with fewer than 300 reads or more than 20,000 reads, were excluded from the gene count matrix. Genes present in fewer than 50 cells were also excluded. Cells with barcode errors (0.4% of total cells) were excluded. Cells with more than 4% mitochondrial reads were removed, and doublet detection was performed using the scanpy.external.pp.scrublet function. The anticipated doublet rate was set to 0.06, and the number of neighbors was configured to 30. Cells with doublet scores exceeding 0.2 were annotated as suspected doublets and excluded from analysis. The data were subsequently normalized and log-transformed. Highly variable genes (5,332 genes) were identified using the scanpy.pp.highly_variable_genes function with parameters min_mean=0.0125, max_mean=3, and min_disp=0.5. The effects of total counts per cell and the proportion of mitochondrial reads per cell were regressed out using the scanpy.pp.regress_out function. The data were then scaled, followed by dimensionality reduction using principal component analysis (PCA) with the scanpy.tl.pca function (svd_solver='arpack'). Batch effects between SHARE-seq batches were corrected using the scanpy.external.pp.harmony_integrate function, with cells stratified by flowcell. A neighborhood graph was computed using the scanpy.pp.neighbors function with 30 neighbors (metric='cosine'). This graph was embedded into two dimensions using the scanpy.tl.umap function, with the minimum effective distance between embedded points set to 0.1. Leiden clustering was carried out using the scanpy.tl.leiden function. For single-cell cluster annotation, a curated list of marker genes mentioned in the original paper[15] was compiled from established cellular reference datasets.

For SHARE-RNA from mixed cell line of HCT116 and NK92 cell lines, cells with fewer than 1,000 genes or more than 6,500 genes detected, as well as those with fewer than 1,000 or more than 20,000 reads, were excluded from the gene count matrix. Additionally, genes present in fewer than 50 cells were removed. The percentage of mitochondrial reads was calculated for each cell and cells with over 30% mitochondrial reads were filtered out. Cell doublets were estimated using the same expected overall doublet rate and the number of neighbors previously. Cells with the doublet score greater than 0.2 were labeled as potential doublets and excluded from further analysis. Following normalization and log-transformation of the data, 7,788 highly variable genes were identified (min_mean = 0.0125, max_mean = 3, min_disp = 0.5). The effects of total counts per cell and mitochondrial read percentage were regressed out. The data were then scaled and dimensionality reduction was performed using PCA with the svd_solver set to 'arpack'. A neighborhood graph of cells was generated using 15 neighbors (metric = 'cosine') and this graph was embedded in two dimensions using uniform manifold approximation and projection (UMAP) with an effective minimum distance of 0.5 between embedded points. Leiden clustering was conducted and differentially expressed genes for each leiden cluster were identified using the scanpy.tl.rank_genes_groups function (method = 'wilcoxon'). These DEGs were subsequently used for cluster annotation.

2.3.3. snATAC-seq data analysis

All snATAC-seq analysis was conducted on Signac. For SHARE-ATAC data from human kidney tissue, all key parameters were followed as outlined in the paper[15]. Chromatin profiling began with the generation of a chromatin assay from the count matrix using the CreateChromatinAssay function in Signac, followed by its conversion into a seurat object using the CreateSeuratObject function from Seurat[33] (v5.1.0). For each cell, nucleosome signal intensity, transcription start site (TSS) enrichment score, fraction of reads in peaks (FRiP), and the proportion of counts overlapping the hg19 genome blacklist were calculated using the NucleosomeSignal, TSSEnrichment, FRiP, and FractionCountsInRegion functions, respectively. Cells were retained if they met the following criteria: 400 to 50,000 peaks, nucleosome signal value below 2.5, TSS enrichment score above 1, FRiP value greater than 0.1, and a blacklist overlap ratio below 0.05. The data were then normalized using the TF-IDF (term frequency-inverse document frequency) method implemented in the RunTFIDF function. Linear dimensionality reduction was achieved through singular value decomposition (SVD) of the TF-IDF matrix using the RunSVD function. Harmony was used to eliminate potential batch effects across cells, stratified by flowcell. Cell clustering, non-linear dimensionality reduction, and UMAP visualization were carried out using the FindNeighbors, FindClusters, and RunUMAP functions, respectively, with parameters set to dims = 2:30, min.dist = 0.1, and n.neighbors = 50. Gene annotation was performed using the GeneActivity function, which computed counts for each cell across gene bodies and 2,000 bp upstream of transcription start sites (including promoter regions) and genes mentioned in the original paper[15].

For SHARE-ATAC data from mixed cell line of HCT116 and NK92 cell lines, a chromatin assay was constructed from the count matrix and subsequently converted into a seurat object. Quality control criteria for cells included having 2,000 to 50,000 peaks, a nucleosome signal value below 2.5, a TSS enrichment score above 4, a FRiP value exceeding 0.1, and a blacklist overlap ratio below 0.05 for the hg38 genome. The data were then normalized and dimensionality reduction was performed. Graph-based clustering, non-linear dimensionality reduction, and UMAP visualization were performed respectively, with parameters dims = 2:30, min.dist = 0.5, and n.neighbors = 30. Gene annotation was performed using the GeneActivity function, which computed counts for each cell across gene bodies and 2,000 bp upstream of transcription start sites (including promoter regions) and DEG lists derived from the SHARE-RNA data.

2.3.4. Integration of snRNA-seq and snATAC-seq across modalities

Cells that met the quality control criteria for both snRNA-seq and snATAC-seq were used for cross-modality integration. After quality control for each modality, the datasets were combined into a seurat object, and dimensionality reduction was performed for each modality following the same procedures as described previously. The Weighted Nearest

Neighbor (WNN) graph[34] was computed using the FindMultiModalNeighbors function, which integrated the dimensionality reduction results from both modalities. For the human kidney data, the following parameters were used: dims.list = list(1:30, 2:30) and k.nn = 30. For the in-house mixed sample data, the following parameters were used: dims.list = list(1:50, 2:50) and k.nn = 20. The WNN graph was then used for UMAP visualization and clustering. For the human kidney data, the parameters min.dist = 0.001 and n.neighbors = 30 were used, while for the in-house mixed sample data, the parameters min.dist = 0.1 and n.neighbors = 50 were applied.

2.4. Linked Peak-gene association in cis chromatin and identification of DORC

To identify peak-gene associations in cis chromatin, FigR[35] (v0.1.0) R package was used. Briefly, FigR calculates Spearman correlation coefficient of each peak-gene pair by considering all peak counts from snATAC-seq located in 100kb window around TSS of each gene and their gene expression values. To estimate the background, chromVAR[36] (v1.26.0) was utilized to generate a null distribution of Spearman correlations between peaks and genes, independent of their peak-gene proximity. It then computes the expected population mean (pop.mean) and standard deviation (pop.sd) from the expected Spearman correlations. The Z score is calculated using the formula $z = (obs-pop.mean)/pop.sd$, and p-values are determined. Only peak-gene associations with the most significant p-values are retained when a peak is linked to multiple genes. To identify a set of proximal peaks for each gene, referred to as DORCs, genes are ranked based on the count of peaks with significant associations (50 kb around TSSs, $p < 0.05$). A cutoff of 5 peaks per gene is applied for in-house SHARE-seq data from a mixed sample of HCT116 and NK92 cell lines. Then peak counts are normalized by the total number of unique fragments in peaks per cell and DORC scores for each gene in each cell were determined by summing the mapped read counts of all significantly correlated peaks per gene, based on a recalculated peak-gene association. The DORC score for each cell type was computed as the average across all cells within that cell type.

2.5. Gene ontology (GO) analysis

GO analysis was conducted using enrichR[37] (v3.2) R package with the GO Biological Process 2023 dataset.

2.6. Super-enhancer calling

Super enhancer regions were identified using ROSE[38,39]. Briefly, this algorithm links adjacent enhancers if they are within 12,500 bp of each other and ranks them based on their H3K27ac signal after subtracting the input signal. It then ranks enhancers in descending order based on H3K27ac signal and identifies super-enhancers as those above the inflection point of the signal. All enhancer regions that overlap with the promoter region (transcription start site ± 2,000 bp) were excluded before stitching.

2.7. Data visualization

Genome tracks were visualized using pyGenomeTracks[40].

2.8. TF-DORC association inference

The inference of TF-DORC associations was performed using the runFigRGRN function by FigR, leveraging the human motif database derived from cisBP. Briefly, this process utilizes the frequency of matches between transcription factor (TF) motifs and peaks to calculate the relative enrichment of TF motifs through the Z-tests. Based on this, the regulation score is defined by combining the significance levels of correlation and peak enrichment, where the correlation is calculated as the Spearman correlation between smoothed DORC accessibility scores and smoothed RNA expression levels across all cells.

2.9. Motif enrichment analysis

Motif enrichment analysis was performed on peak regions associated with DORC genes using findMotifsGenome.pl in HOMER[41]. The known motif analysis was utilized for data interpretation.

3. Results

3.1. Validation of reproducibility in cross-modality integrated analysis pipeline for single-cell transcriptomic and epigenomic data from human kidney tissue

To evaluate the reproducibility and reliability of a bioinformatic pipeline designed for the simultaneous integrated analysis of single-nucleus RNA sequencing (snRNA-seq) and ATAC sequencing (snATAC-seq) from the same cells (**Figure 1A**), the SHARE-seq public data from Li et al.[15], profiling transcriptomic and epigenomic landscapes at single-cell resolution across diverse anatomical regions of human kidney tissue were employed. In the upstream processing steps, barcode demultiplexing, alignment to the reference genome, and the generation of a single-cell-level gene expression and fragment-per-cell count matrices for downstream analysis were performed (**Tables 1, 2**). For the snRNA-seq data, the count matrix was processed by filtering out cells with abnormal gene expression, followed by log normalization and batch correction to minimize technical biases and enhance the comparability of biologically derived expression differences (**Figure 1B**). To ensure clear separation of cell type clusters and reduce noise, highly variable genes (n = 5,322) and principal components (PCs) that explained the major sources of variability were selected (**Figure 1C**). Clustering was then applied to distinguish similar from dissimilar ones, and cell type annotation was conducted using marker gene expression patterns. For the snATAC-seq data, peaks were identified from the fragments-per-cell information as regions with high chromatin accessibility and converted into a peak-by-cell count matrix. Cells with abnormal accessibility patterns were filtered (**Figure 1D**), followed by TF-IDF normalization, batch correction, selection of key latent semantic indexing (LSI) components, dimensionality reduction, and clustering (**Figure 1B**). Cell type annotation was subsequently performed based on marker gene activity profiles of each cell type. The integration of snRNA-seq and snATAC-seq across modalities was performed using the intersected barcodes shared between the two modalities, incorporating both gene expression and chromatin accessibility information (**Figure 1B**). The process included data normalization, dimensional reduction, and batch correction, which were re-performed for both modalities. Using the results of dimensionality reduction from each modality, a weighted nearest neighbor method was applied (**Figure 1B**). The WNN graph was constructed using the principal components (dimensions 1–30) from RNA and the latent semantic indexing components (dimensions 2–30) from ATAC, which were considered to explain the primary sources of variation.

Our analysis successfully reproduced the transcriptional and epigenetic UMAP representations of human kidney tissue, identifying 29 distinct clusters for snRNA-seq and 21 for snATAC-seq (**Figures 2A, 2C**). These results revealed distinct clustering patterns that closely matched those reported in the original study. Additionally, the gene expression profiles were faithfully reproduced, with clear and distinct expression patterns observed for marker genes representing each cell type (**Figure 2B**). Joint analysis of both modalities using the WNN method on cells with shared barcodes (n = 324,701), as mentioned in the original paper, defined precise cellular states based on multiple data types and revealed similar patterns of cell-type-specific regulatory features and marker gene expression, effectively reflecting the characteristics observed in the individual modality analyses (**Figure 2D**).

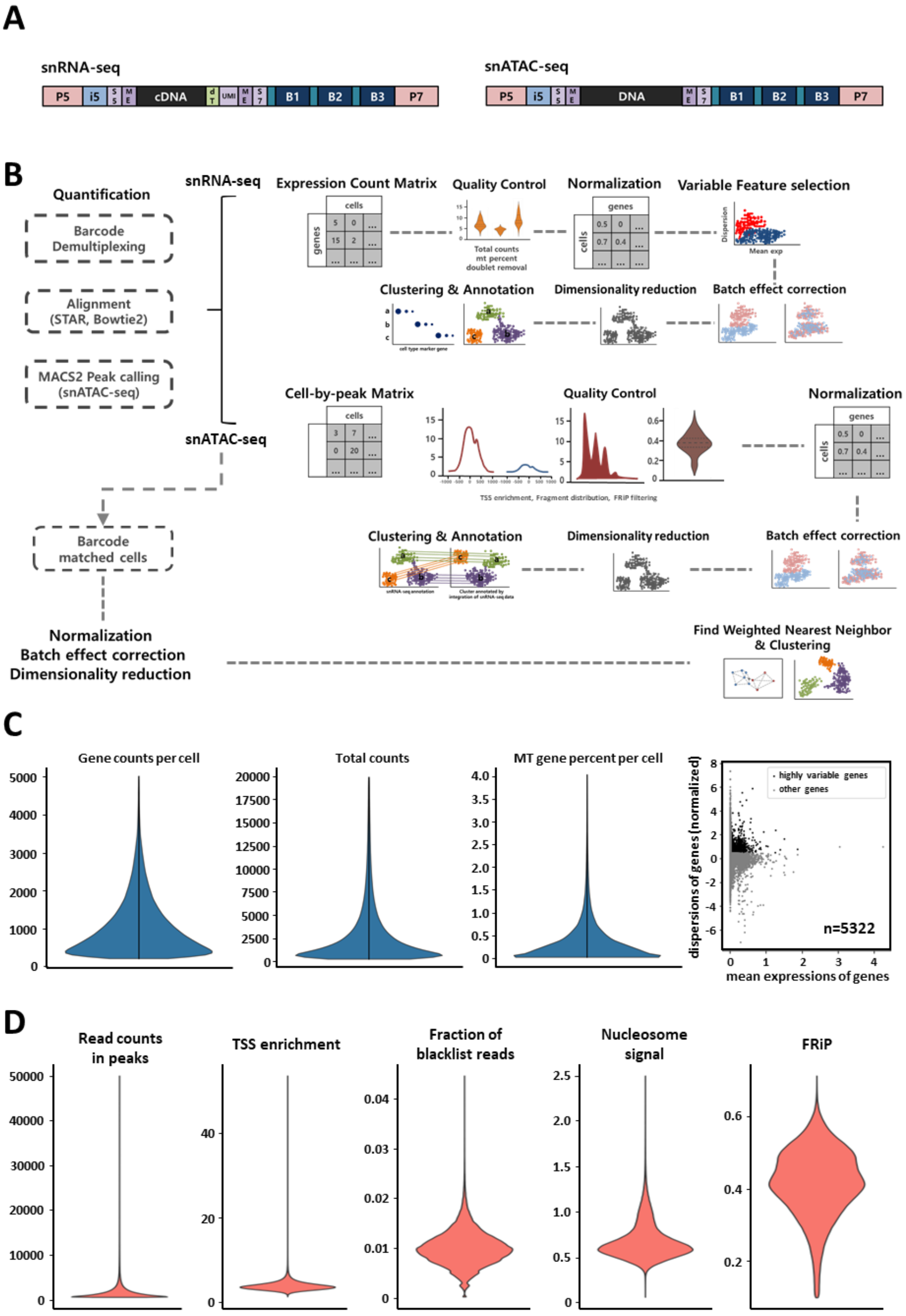
A
snRNA-seq
P5
i5
S5
ME
cDNA
dT
UMI
ME
S7
B1
B2
B3
P7
snATAC-seq
P5
i5
S5
ME
DNA
ME
S7
B1
B2
B3
P7
B
Quantification
Barcode Demultiplexing
Alignment (STAR, Bowtie2)
MACS2 Peak calling (snATAC-seq)
snRNA-seq
Expression Count Matrix
cells
genes
Quality Control
Total counts
mt percent
doublet removal
Normalization
genes
cells
Variable Feature selection
Dispersion
Mean exp
Clustering & Annotation
cell type marker gene
Dimensionality reduction
Batch effect correction
snATAC-seq
Cell-by-peak Matrix
cells
Quality Control
TSS enrichment, Fragment distribution, FRiP filtering
Normalization
genes
cells
Clustering & Annotation
snRNA-seq annotation
Cluster annotated by integration of snRNA-seq data
Dimensionality reduction
Batch effect correction
Barcode matched cells
Normalization
Batch effect correction
Dimensionality reduction
Find Weighted Nearest Neighbor & Clustering
C
Gene counts per cell
Total counts
MT gene percent per cell
highly variable genes
other genes
n=5322
dispersions of genes (normalized)
mean expressions of genes
D
Read counts in peaks
TSS enrichment
Fraction of blacklist reads
Nucleosome signal
FRiP

Figure 1. Overview of SHARE-seq data analysis from human kidney tissue. (A) The structure of snRNA-seq and snATAC-seq libraries of SHARE-seq. (B) The upstream and downstream analysis processing workflow of SHARE-seq data. (C) The violin plot illustrates the filtering results from downstream QC performed on snRNA-seq data from human kidney tissue, alongside the results of highly variable gene selection. (D) The violin plot illustrates the filtering results from downstream QC performed on snATAC-seq data from human kidney tissue.

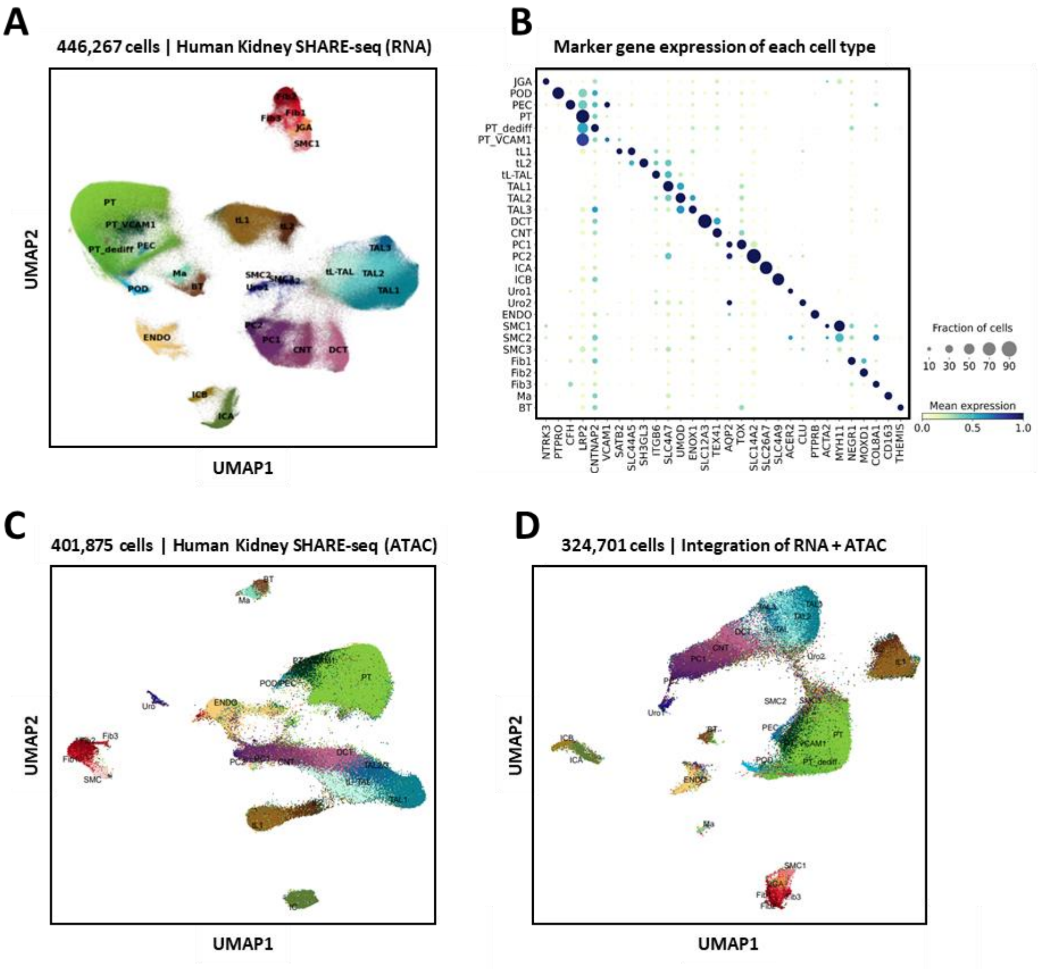
A
446,267 cells | Human Kidney SHARE-seq (RNA)
UMAP2
UMAP1
PT
PT_VCAM1
PT_dediff
PEC
POD
Ma
BT
ENDO
tL1
tL2
SMC2
Uro1
tL-TAL
TAL3
TAL2
TAL1
PC2
PC1
CNT
DCT
ICB
ICA
Fib3
Fib1
JGA
SMC1
B
Marker gene expression of each cell type
JGA
POD
PEC
PT
PT_dediff
PT_VCAM1
tL1
tL2
tL-TAL
TAL1
TAL2
TAL3
DCT
CNT
PC1
PC2
ICA
ICB
Uro1
Uro2
ENDO
SMC1
SMC2
SMC3
Fib1
Fib2
Fib3
Ma
BT
NTRK3
PTPRO
CFH
LRP2
CNTNAP2
VCAM1
SATB2
SLC44A5
SH3GL3
ITGB6
SLC4A7
UMOD
ENOX1
SLC12A3
TEX41
AQP2
TOX
SLC14A2
SLC26A7
SLC4A9
ACER2
CLU
PTPRB
ACTA2
MYH11
NEGR1
MOXD1
COL8A1
CD163
THEMIS
Fraction of cells
10 30 50 70 90
Mean expression
0.0 0.5 1.0
C
401,875 cells | Human Kidney SHARE-seq (ATAC)
UMAP2
UMAP1
BT
Ma
PT
POD
PEC
ENDO
Uro
Fib3
SMC
DCT
PC2
CNT
TAL2
tL-TAL
TAL1
D
324,701 cells | Integration of RNA + ATAC
UMAP2
UMAP1
TAL3
TAL2
DCT
CNT
PC1
Uro2
Uro1
SMC2
PEC
PT
PT_VCAM1
POD
PT_dediff
ICB
ICA
BT
ENDO
Ma
SMC1
Fib3

Figure 2. Reproducing the cellular heterogeneity of the human kidney tissue was achieved through SHARE-seq multiomics analysis. (A, C) The UMAP visualization displays (A) 446,267 single-cell transcriptomes and (C) 401,875 single-cell chromatin ccessibility profiles. (B) The dot plot highlights the marker gene expression specific to each cluster. (D) An integrative analysis of both modalities was conducted using WNN on 324,701 intersected cells.

Table 1. Upstream QC for snRNA-seq from human kidney tissue

flowcells	**Uniquely mapped reads (%)**	**mapped to multiple loci (%)**	**Duplication Rate (%)**	**Number of barcodes**
S2	65.37%	32.42%	67.35%	582,354
S4-1	61.07%	36.15%	43.11%	
S4-2	68.78%	28.65%	64.49%	

Table 2. Upstream QC for snATAC-seq from human kidney tissue

flowcells	**Alignment Rate (%)**	**Duplication Rate (%)**	**MT Rate (%)**	**TSS Rate**	**Number of barcodes**
S2	98.69%	40.37%	7.7%	11.79	970,490
S4-1	98.72%	40.68%	8.17%	12.08	
S4-2	98.59%	51.05%	2.28%	7.68	

3.2. Distinct transcriptional and chromatin accessibility profiles of each cell type revealed by single-cell analysis of in-house SHARE-seq data from a mixed sample of HCT116 and NK92 cell lines

Building upon the successful reproduction of integrative analysis from the public SHARE-seq data, we utilized the same pipeline on in-house SHARE-seq data from a mixed sample of two distinct cell lines: HCT116 (human colorectal carcinoma) and NK92 (natural killer cells). These cell lines exhibit markedly different transcriptional and chromatin accessibility profiles, making it a biologically heterogeneous sample. First, to ensure the reliability of the analysis results, we performed upstream QC using the same process as mentioned above (**Figures 3A, 3B; Tables 3, 4**). The results showed that 75.45% of the snRNA-seq library consisted of uniquely mapped reads, while 23.13% were multimapped reads, indicating a high mapping rate to the human genome reference (hg38). The duplication rate was approximately 46%, and a total of 736,448 barcodes were identified. This number is much higher than the 20,000 cells used in library preparation, a phenomenon observed similarly in the snATAC-seq data. This discrepancy is likely due to the characteristics of SHARE-seq's split-pool barcoding method, which uses around 1,000,000 barcodes (96*96*96), leading to barcode assignment to molecules not originating from the nucleus, among other factors. In terms of functional distribution, the majority of reads (~41%) mapped to intronic regions, reflecting the capture of nuclear RNA such as pre-mRNA in the splicing intermediate state. Other reads mapped to intergenic regions (~30%), protein coding sequences (CDS, ~20%), and untranslated regions (UTR, ~10%), and they show high signal in specific marker gene expression of each cell type (**Figure 3C**). For the snATAC-seq library, the alignment rate was 98.59%, with a relatively low duplication rate of 17.35%, indicating a satisfactory level of library complexity. The mitochondrial read ratio was 4.96%, suggesting minimal contamination by extrachromosomal DNA, and the TSS enrichment score was 21.76, indicating a strong signal in active promoter regions. The insert size distribution, as expected in ATAC-seq, showed the highest percentage of fragments corresponding to the nucleosome-free region (NFR), with distinct regions in the areas between the NFR and 200 bp, 400 bp, and 600 bp, corresponding to nucleosome-bound DNA fragments. The signal decreased progressively in these regions, highlighting the strong chromatin accessibility at transcriptionally active sites and a high signal-to-noise ratio (**Figure 3C**).

Table 3. Upstream QC for snRNA-seq from a mixed sample of HCT116 and NK92

Uniquely mapped reads (%)	**mapped to multiple loci (%)**	**Duplication Rate (%)**	**Number of barcodes**
75.45 %	23.13 %	46.14 %	736,448

Table 4. Upstream QC for snATAC-seq from a mixed sample of HCT116 and NK92

Alignment Rate (%)	**Duplication Rate (%)**	**MT Rate (%)**	**TSS Rate**	**Number of barcodes**
98.59 %	17.35 %	4.96 %	21.76	884,378

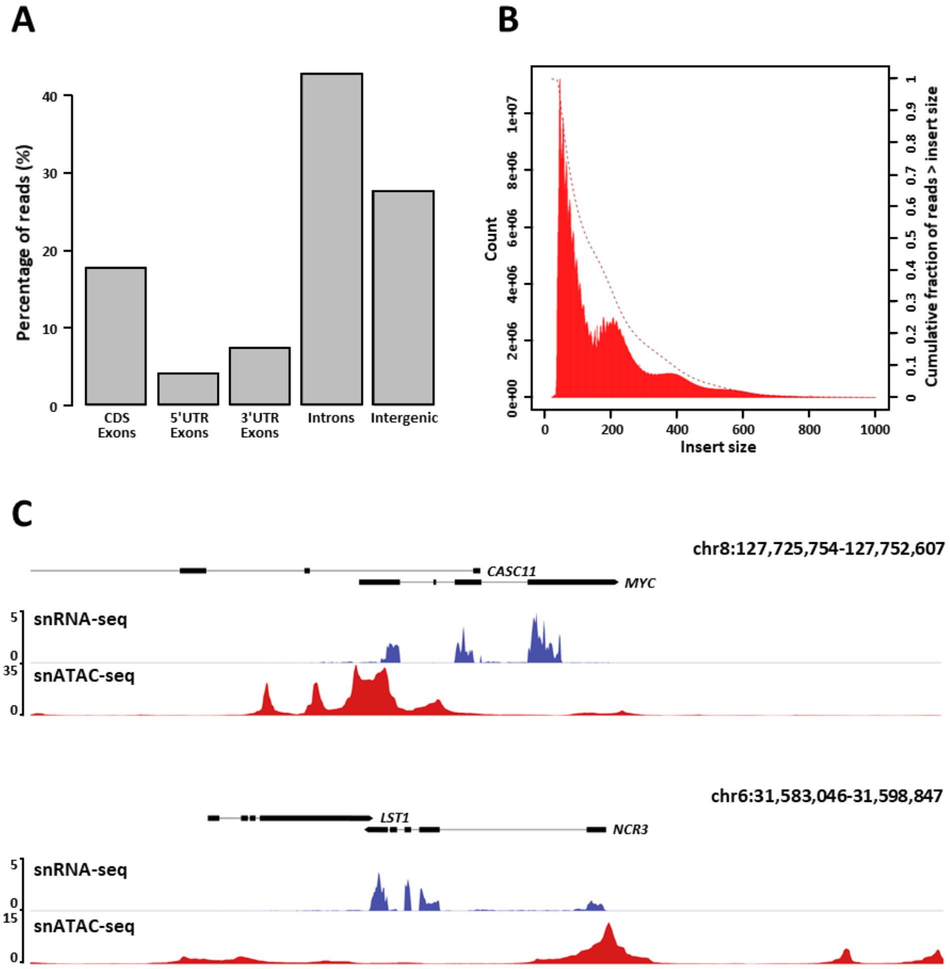

Figure 3. The upstream QC results from in-house SHARE-seq data from a mixed sample of HCT116 and NK92 cell lines. (A) Read distribution of snRNA-seq from a mixed sample of HCT116 and NK92 cell lines. (B) Insert size histogram for all reads of snATAC-seq from a mixed sample of HCT116 and NK92 cell lines. (C) Gene expression and chromatin accessibility profiles of cell-type specific marker genes.

In the downstream QC of the snRNA-seq gene expression per cell count matrix, excessive barcode detection resulted in notably low total counts and expressed gene numbers in most cells. Furthermore, a distinct barcode distribution pattern was identified, with cells showing very low counts clustered in one region and another group forming a secondary cluster slightly above this region (**Figure 4A**). Cells with appropriate levels of expression and read counts in this region were considered as derived from the 20,000 cells initially used, and cells exhibiting abnormally low or high read counts or expressed gene numbers were filtered out. The mitochondrial rate for each cell was calculated, and cells with excessively high mitochondrial gene expression were removed. Potential doublets were identified and removed by filtering out cells with high scores indicating the presence of the same barcode in two cells, resulting in a final count of 14,610 cells (**Figures 4B, 4C**) and 7,788 highly variable genes (**Figure 4D**). Following dimensionality reduction via PCA, three main approaches were applied to identify the appropriate number of principal components (PCs) for clustering. The first method involved selecting PCs based on cumulative variance, considering the point where the cumulative variance exceeded 90% and excluding PCs where the individual variance explained was less than 1%. The second approach determined the number of PCs by identifying the point where the difference in variance between consecutive PCs exceeded 0.1%. Finally, the third method utilized visual assessment through a Scree plot, excluding PCs beyond the point where the variance explained showed a sharp decline. As the number of selected PCs increases, more variance can naturally be explained. However, beyond a certain point, overfitting may occur, potentially introducing noise into the data. From this perspective, it was determined that PCs selected solely through computational methods do not always yield optimal results in reflecting the characteristics of the cell types on clustering. Specifically, PCs chosen based on cumulative variance or variance differences between consecutive PCs were found to be prone to overfitting or insufficient for clearly distinguishing subtle differences or biological features between the two cell types. Taking all three approaches into account, the optimal number of PCs was determined to be 5, as this corresponded to the elbow point in the Scree plot, where the variance explained by additional PCs became negligible (**Figure 4E**). Subsequently, UMAP dimensionality reduction was applied to evaluate the separation between cell types. The number of neighbors in UMAP was tested across a range of values, and the dimensionality reduction results were compared across different combinations of PC numbers in the Scree plot above and neighbor sizes (**Figure 5A**). Based on the UMAP results for various combinations of PCs and neighbor sizes, the combination of PC 5 and neighbor 15 was most likely to effectively distinguish the two cell types and accurately capture their characteristic features. This process enabled the identification of the optimal combination of PC number and neighbor size, which resulted in the delineation of three clusters, thereby optimizing cell type separation. Analysis of the differentially expressed genes (DEGs) in each cluster revealed that Cluster 0, in particular, contains a significant number of genes regulating biological processes closely associated with cancer cell characteristics (**Figure 6A**). Gene Ontology analysis showed that the DEGs in this cluster are significantly enriched in processes such as regulation of epithelial cell proliferation, regulation of epidermal growth factor receptor activity, regulation of cell migration, angiogenesis, and the ERBB2-EGFR signaling pathway, all of which are hallmark features of cancer. These findings suggest that this cluster exhibits properties resembling those of cancer cells, including invasiveness, metastatic potential, enhanced proliferation, and modulation of the tumor microenvironment, consistent with the characteristics of the HCT116 cell line (**Figure 5B**). Cluster 1 contains a significant number of genes involved in several key biological processes related to natural killer (NK) cells (**Figure 6A**). Gene Ontology (GO) analysis revealed that the DEGs in this cluster are significantly associated with immune processes, such as antigen receptor-mediated signaling pathway, positive regulation of cytokine production, cellular defense response, regulation of interleukin-2 production. Notably, GO terms related to regulation of natural killer cell mediated cytotoxicity and positive regulation of natural killer cell mediated immunity were prominently represented, suggesting that this cluster exhibits characteristics similar to those of NK cell lines, reflecting immune cell-like properties. On the other hand, for Cluster 2, no significant GO terms were identified among the DEGs. The DEGs of cluster 2 show a pattern that is more similar to cluster 0, which is predicted to be the HCT116 cell line, rather than cluster 1, which is predicted to be NK92 cells. However, the overall expression profile of cluster 2 appears to be characterized by a high expression of ribosomal RNA genes (**Figure 6B**). This suggests that the influx of rRNA into the HCT116 cell line during the snRNA-seq experiment may have influenced the gene expression data of these cells. The expression of well-known marker genes for both HCT116 and NK92 cells was found to be divided into two distinct patterns across the clusters. Clusters 0 and 2 showed similar marker gene expression patterns to the HCT116 cell line, while cluster 1 exhibited a clear expression pattern characteristic of the NK92 cell line (**Figures 6C–6E**).

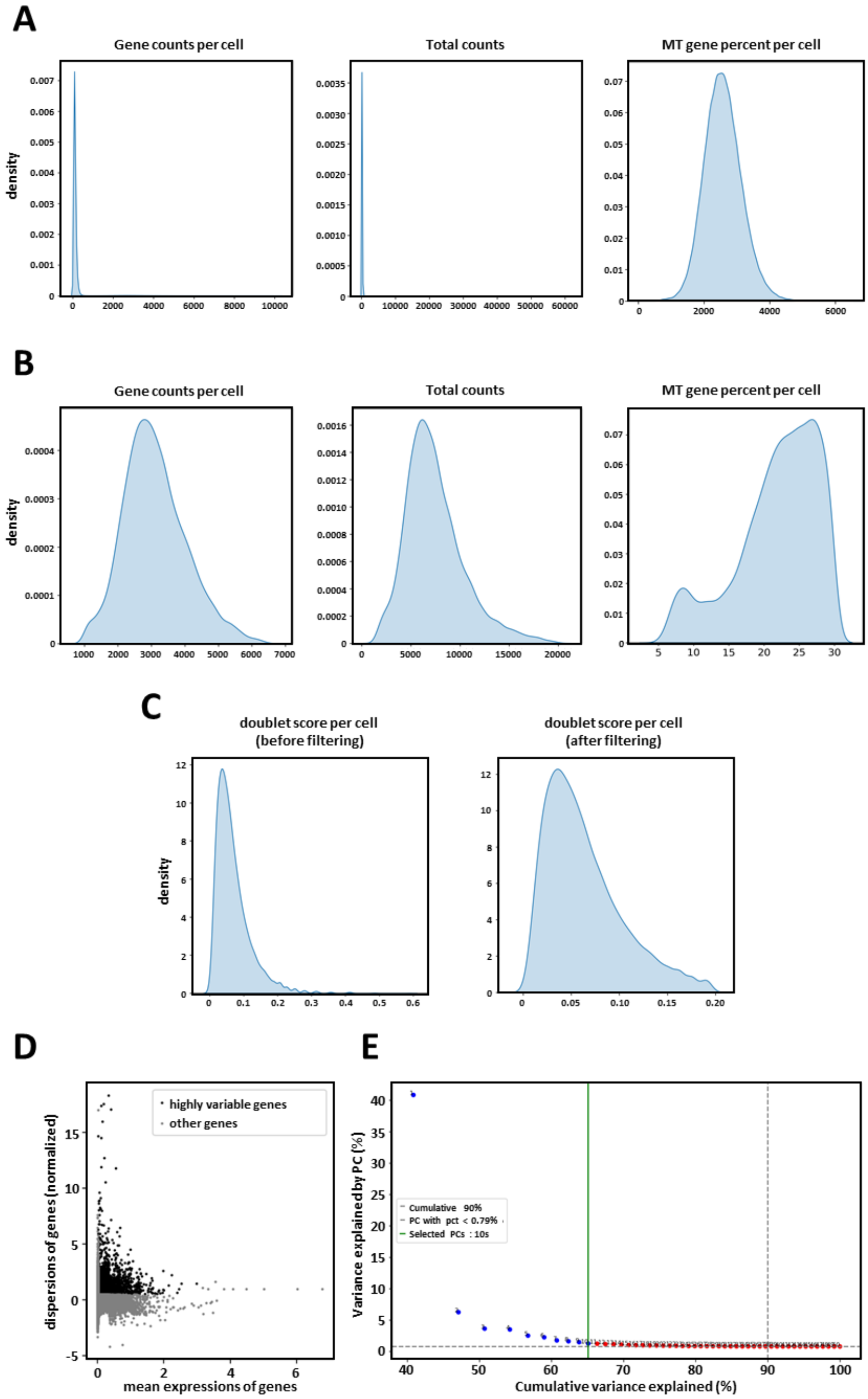
A
Gene counts per cell
density
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
0 2000 4000 6000 8000 10000
Total counts
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0
0 10000 20000 30000 40000 50000 60000
MT gene percent per cell
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 2000 4000 6000
B
Gene counts per cell
density
0.0004
0.0003
0.0002
0.0001
0
1000 2000 3000 4000 5000 6000 7000
Total counts
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0
0 5000 10000 15000 20000
MT gene percent per cell
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
5 10 15 20 25 30
C
doublet score per cell
(before filtering)
density
12
10
8
6
4
2
0
0 0.1 0.2 0.3 0.4 0.5 0.6
doublet score per cell
(after filtering)
12
10
8
6
4
2
0
0 0.05 0.10 0.15 0.20
D
highly variable genes
other genes
dispersions of genes (normalized)
15
10
5
0
-5
0 2 4 6
mean expressions of genes
E
Cumulative 90%
PC with pct < 0.79%
Selected PCs : 10s
Variance explained by PC (%)
40
35
30
25
20
15
10
5
0
40 50 60 70 80 90 100
Cumulative variance explained (%)

Figure 4. Downstream processing results of in-house snRNA-seq data from a mixed sample of HCT116 and NK92 cell lines. (A) The QC distribution patterns of snRNA-seq data from a mixed sample of HCT116 and NK92 cell lines. The overall trends are depicted by the density plot. (B-C) The first QC results of snRNA-seq data from a mixed sample of HCT116 and NK92 cell lines. (D) The distribution of dispersions for 7,788 highly variable genes. Highly variable genes are represented as black dots. (E) A Scree plot presenting the cumulative variance accounted for by each principal component. The number of principal components that explain more than 90% of the variance is indicated by a gray dashed line, while the point where the variance change becomes negligible is marked by a green line.

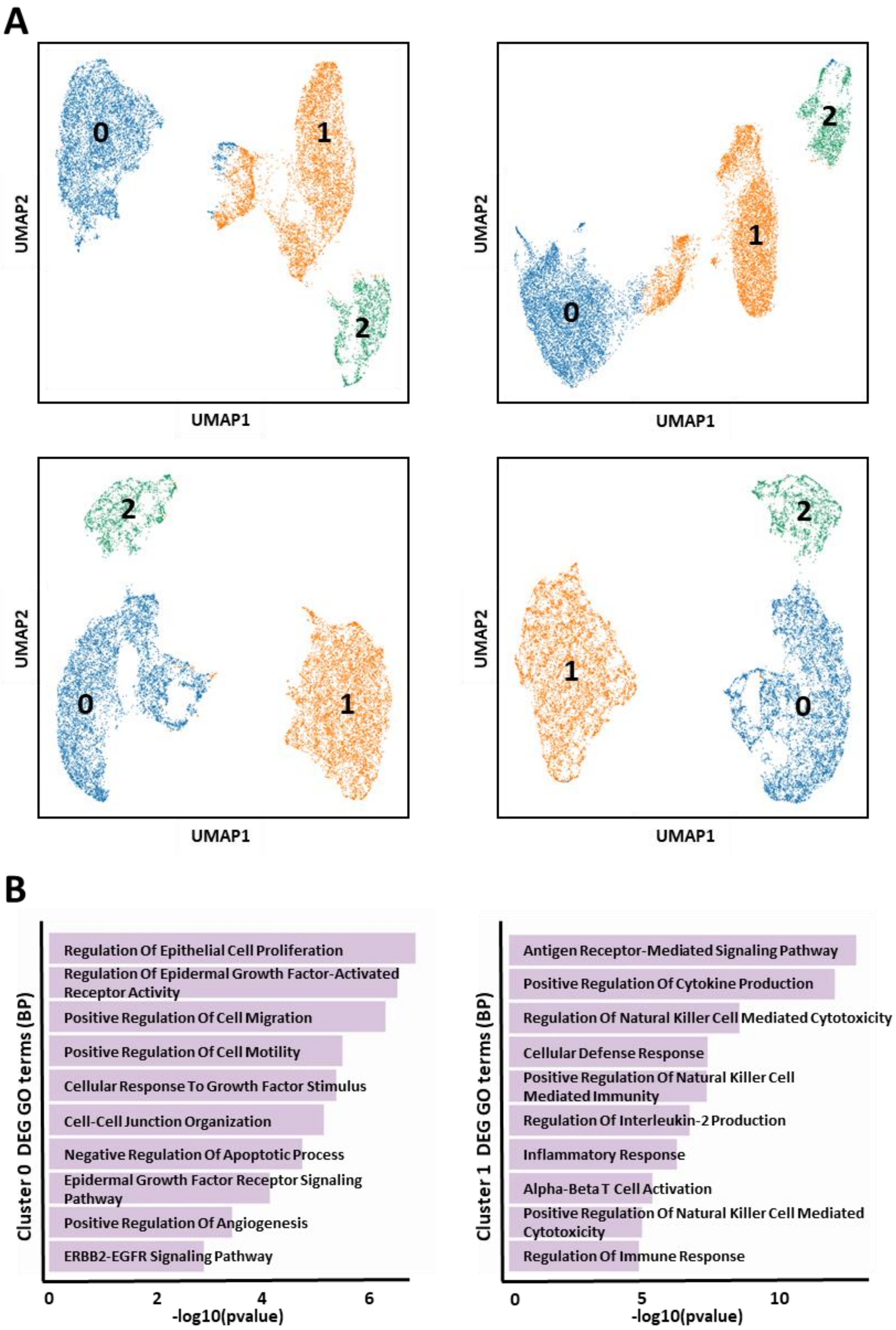
A
0
1
2
UMAP2
UMAP1
2
1
0
UMAP2
UMAP1
2
0
1
UMAP2
UMAP1
2
1
0
UMAP2
UMAP1
B
Cluster 0 DEG GO terms (BP)
Regulation Of Epithelial Cell Proliferation
Regulation Of Epidermal Growth Factor-Activated Receptor Activity
Positive Regulation Of Cell Migration
Positive Regulation Of Cell Motility
Cellular Response To Growth Factor Stimulus
Cell-Cell Junction Organization
Negative Regulation Of Apoptotic Process
Epidermal Growth Factor Receptor Signaling Pathway
Positive Regulation Of Angiogenesis
ERBB2-EGFR Signaling Pathway
0
2
4
6
-log10(pvalue)
Cluster 1 DEG GO terms (BP)
Antigen Receptor-Mediated Signaling Pathway
Positive Regulation Of Cytokine Production
Regulation Of Natural Killer Cell Mediated Cytotoxicity
Cellular Defense Response
Positive Regulation Of Natural Killer Cell Mediated Immunity
Regulation Of Interleukin-2 Production
Inflammatory Response
Alpha-Beta T Cell Activation
Positive Regulation Of Natural Killer Cell Mediated Cytotoxicity
Regulation Of Immune Response
0
5
10
-log10(pvalue)

Figure 5. Cell embedding results of in-house SHARE-seq data from a mixed sample of HCT116 and NK92 cell lines reveal cell type characteristics in each cluster. (A) Cell embedding of in-house SHARE-seq data from a mixed sample of HCT116 and NK92 cell lines. The UMAPs, shown in order from the top left, use PC10, PC38, PC7, and PC5, respectively, with a consistent neighbor parameter of 15 across all plots. (B) Identification of GO terms from biological process associated with differentially expressed genes in clusters 0 and 1.

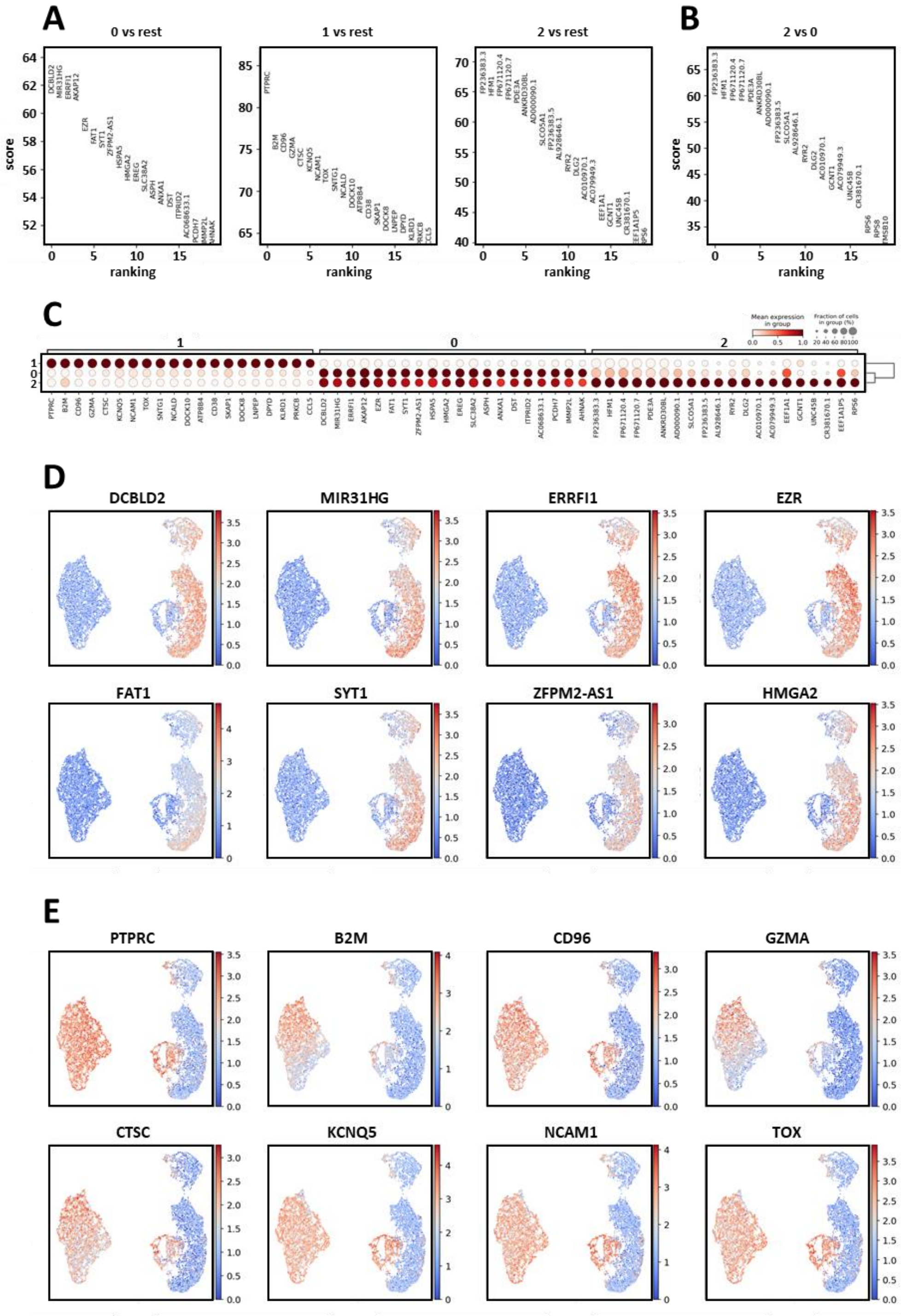
A
0 vs rest
score
ranking
DCBLD2
MIR31HG
ERRFI1
AKAP12
EZR
FAT1
SYT1
ZFPM2-AS1
HSPA5
HMGA2
EREG
SLC38A2
ASPH
ANXA1
DST
ITPRID2
AC068633.1
PCDH7
IMMP2L
AHNAK
1 vs rest
PTPRC
B2M
CD96
GZMA
CTSC
KCNQ5
NCAM1
TOX
SNTG1
NCALD
DOCK10
ATP8B4
CD38
SKAP1
DOCK8
LNPEP
DPYD
KLRD1
PRKCB
CCL5
2 vs rest
FP236383.3
HFM1
FP671120.4
FP671120.7
PDE3A
ANKRD30BL
AD000090.1
SLCO5A1
FP236383.5
AL928646.1
RYR2
DLG2
AC010970.1
AC079949.3
EEF1A1
GCNT1
UNC45B
CR381670.1
EEF1A1P5
RPS6
B
2 vs 0
RPS8
TMSB10
C
Mean expression in group
0.0 0.5 1.0
Fraction of cells in group (%)
20 40 60 80 100
1
0
2
D
DCBLD2
MIR31HG
ERRFI1
EZR
FAT1
SYT1
ZFPM2-AS1
HMGA2
E
PTPRC
B2M
CD96
GZMA
CTSC
KCNQ5
NCAM1
TOX

Figure 6. Cluster-specific differential gene expression and UMAP visualization in SHARE-seq data from a mixed sample of HCT116 and NK92 cell lines. (A, C) Differentially expressed genes of each cluster and average expression values. (B) The differentially expressed genes of cluster 2 compared to cluster 0, inferred to be associated with HCT116. (D) The UMAP visualization of each DEG's expression patterns for cluster 0, which is inferred to be associated with HCT116. (E) The UMAP visualization of each DEG's expression patterns for cluster 1, inferred to be associated with NK92.

In the downstream QC of the snATAC-seq accessibility per cell count matrix, as mentioned in the snRNA-seq analysis, an excessive number of barcodes were identified. This led to a notable decrease in fragment counts within peaks from cells that were not derived from normal cells across most barcodes. Cells with appropriate levels of fragment counts within accessible regions were considered as originating from the 20,000 cells initially used, and those with abnormally low or high counts were filtered out. Additionally, cells with excessively low enrichment ratios of reads at TSS or low FRiP values (**Figure 7A**) were removed. Cells with high fractions of reads in blacklist regions or excessive nucleosome signal, which resulted in a lack of enrichment in the NFR, were also excluded from the analysis. After reducing the dimensionality of the data using latent semantic indexing, an appropriate number of LSI components (LSI30) was selected to sufficiently explain the variability in the data while avoiding overfitting. This selection was based on the Scree plot and commonly used metrics. The correlation between sequencing depth and each LSI component was also assessed. For LSI1, the correlation coefficient was nearly 1, indicating that LSI1 primarily explained variability related to technical sequencing depth rather than accessibility patterns (**Figure 7B**). As a result, LSI1 was excluded from further analysis. UMAP dimensionality reduction was subsequently applied to evaluate the separation between cell types. By comparing UMAP results across various combinations of LSI numbers and neighbor sizes, the combination of LSI30 and neighbor 20 was determined to most effectively distinguish between the two cell types and accurately capture their characteristics (**Figure 7C**). Gene activity within accessible regions for each cluster was assessed in the ATAC-seq data. In Cluster 0, gene activity of differentially expressed genes in snRNA-seq associated with cancer cell characteristics were enriched. This suggests that cells in Cluster 0 are accessible at genes that exhibit properties characteristic of cancer cells, being likely to correspond to the HCT116 cell line (**Figure 8A**). In Cluster 1, gene activity of DEGs in snRNA-seq associated with immune cell characteristics was enriched. This suggests that cells in Cluster 1 are accessible at genes that exhibit properties resembling those of natural killer (NK) cells, reflecting key immune functions (**Figure 8B**).

The analysis of the in-house SHARE-seq data from the mixed sample of HCT116 and NK92 cell lines revealed distinct chromatin accessibility patterns and gene expression profiles between the two cell types. Gene activity and expression patterns of each cell type showed high similarity. These findings underscore the potential of the SHARE-seq method in capturing the complexity of heterogeneous cell populations and demonstrate its utility in studying the molecular signatures of distinct cell types.

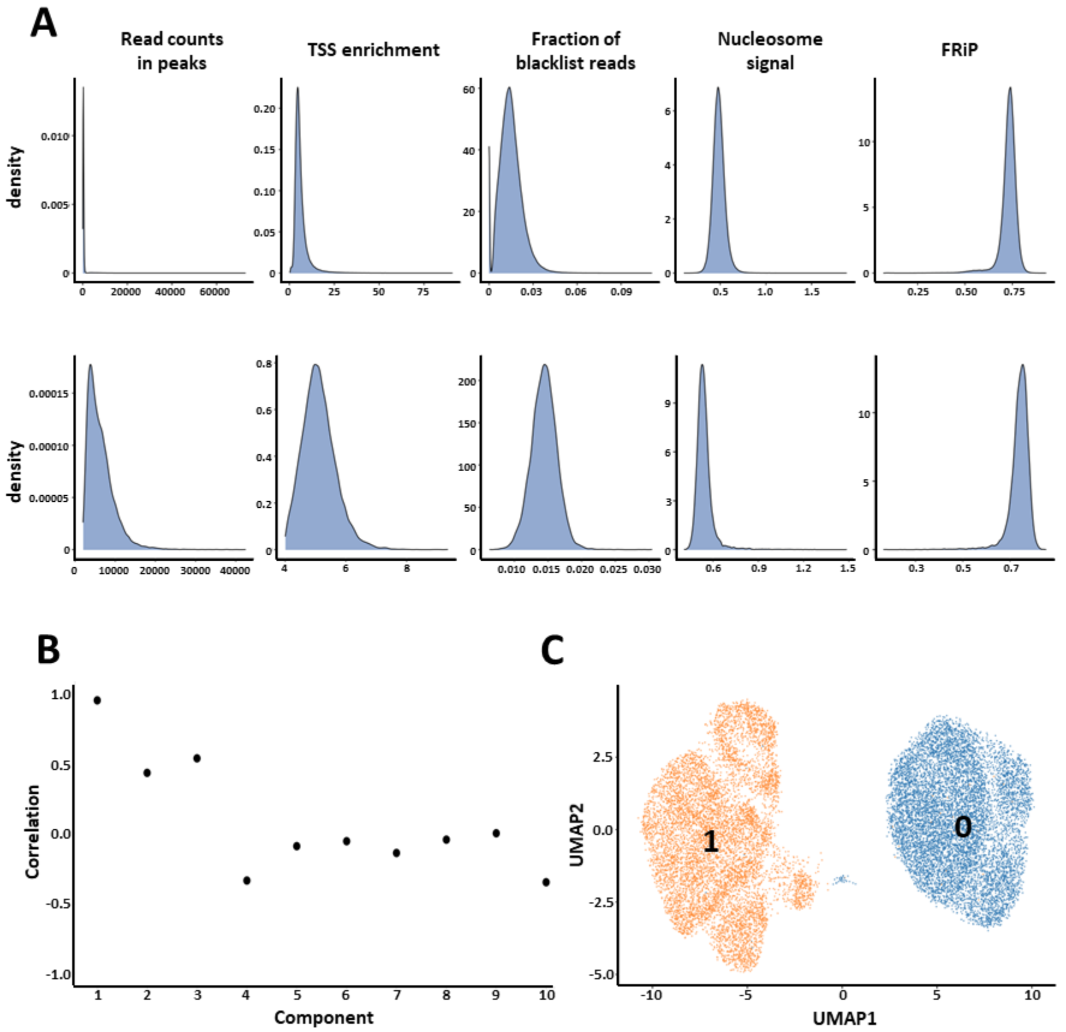

Figure 7. Downstream processing and cell embedding results of in-house SHARE-ATAC data from a mixed sample of HCT116 and NK92 cell lines. (A) The distribution patterns of QC features for snATAC-seq data from the mixed sample of HCT116 and NK92 cell lines, shown before (top) and after (bottom) filtering. The overall trends are depicted by the density plot. (B) A correlation plot between sequencing depth and individual LSI components. (C) The UMAP-based cell embedding results of snATAC-seq data.

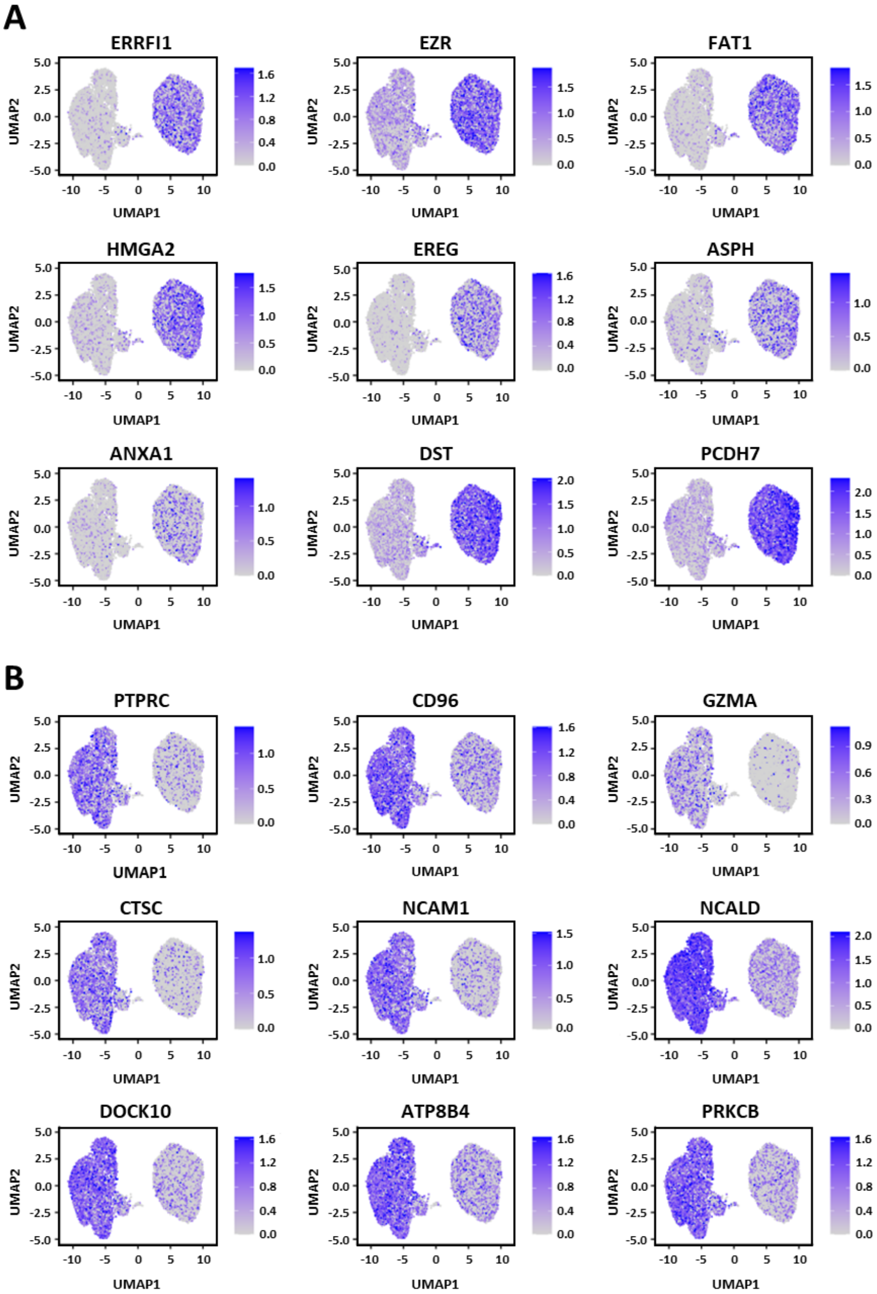

Figure 8. Gene activity of cell type-specific DEGs in in-house snATAC-seq data. (A) Patterns of gene activity that exhibit differential accessibility across clusters.

3.3. Integration of snRNA-seq and snATAC-seq for enhanced cell-type annotation and characterization of cellular heterogeneity

Cells that passed the quality control criteria for both snRNA-seq and snATAC-seq were subsequently analyzed. A total of 14,007 barcodes were matched, corresponding to approximately 95% of the snRNA-seq cells and 80% of the snATAC-seq cells. Integrative analysis of both modalities (snRNA-seq and snATAC-seq) by WNN on the 14,007 cells revealed that cell clustering based on both snRNA-seq and snATAC-seq data largely reflected similar patterns in the cell embeddings (**Figures 9A, 9B**). However, some cells, which were grouped into the same cluster in the snRNA-seq analysis, were separated into distinct clusters upon integration, highlighting differences in cell embedding based on both modalities. For instance, in the snRNA-seq clustering results, among the cells in cluster 0, which was identified as HCT116, the cells that showed differences in marker gene expression patterns within cluster 0 (**Figure 6D**) exhibited a shift in embedding towards a cluster displaying the marker gene expression patterns of NK92 in the WNN clustering (**Figure 9A**). This observation suggests that integrating multimodal data enhances the accuracy of identifying complex differences between cell clusters and improves cell type annotation. To further refine cell typing, clustering was performed based on the WNN embedding results (**Figure 9C**), and these results were compared with the snRNA-seq and snATAC-seq cluster annotations. Cells showing consistent characteristics across cluster annotations from both modalities were classified as HCT116 and NK92 cells, comprising 3,937 cells and 6,036 cells, respectively (**Figure 9D**). The remaining cells were classified as Unknown1 or Unknown2, representing states that were somewhat different from the typical characteristics of the two cell types. These results demonstrate the effectiveness of integrating data from both modalities in improving cell-type annotation accuracy and uncovering subtle cellular heterogeneity. The enhanced resolution of gene expression and chromatin accessibility patterns could be valuable for exploring complex regulatory landscapes and understanding cell-specific functions in various biological contexts.

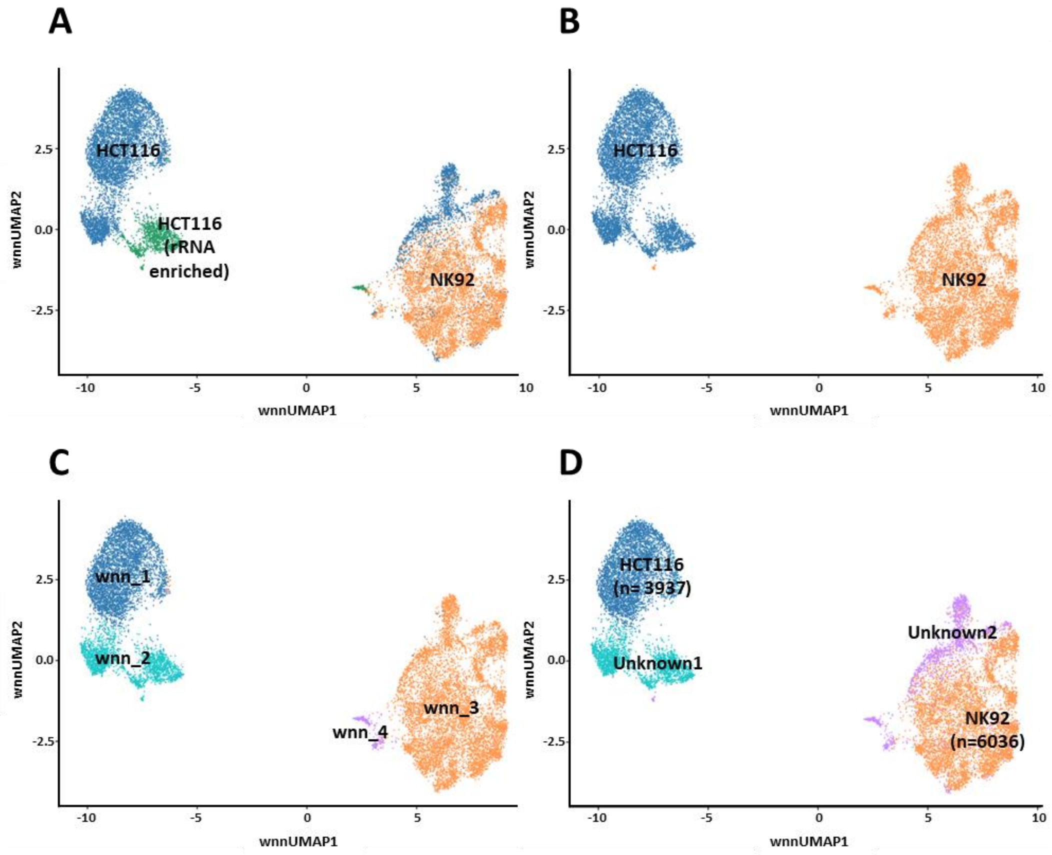

Figure 9. Cross-modality integration of snRNA-seq and snATAC-seq. Integrated cell embedding results of snRNA-seq and snATAC-seq data from a mixed sample of HCT116 and NK92 cell lines using WNN analysis. (A) Cell type annotation was performed based on snRNA-seq clusters. (B) Cell type annotation was performed based on snATAC-seq clusters. (C) Cell type annotation was performed based on clusters defined by the WNN graph. (D) Final cell type annotation from the integrated analysis of snRNA-seq and snATAC-seq data from SHARE-seq of mixed sample (HCT116 & NK92).

3.4. Linking gene expression and chromatin accessibility through SHARE-seq data in cell-type specific domains of regulatory chromatin (DORCs) region

Based on the correlation between peak accessibility near genes in snATAC-seq, which is considered to have a high potential for physical accessibility and therefore the ability to regulate gene expression, and gene expression in snRNA-seq, the cell type-specific cis-regulatory landscape was examined in the two cell types. From the 14,007 cells commonly derived from snRNA-seq and snATAC-seq data, selected HCT116 cells (n = 3,937) and NK92 cells (n = 6,036), which clearly exhibited the characteristics of each cell line (**Figure 9D**), were used to define domains of regulatory chromatin (DORCs) that would highlight the distinct identities of these two cell types. As a result, 76 DORC genes in the HCT116 cell line and 133 DORC genes in the NK92 cell line were identified (**Figures 10A, 10B**). GO analysis revealed that DORC genes identified in HCT116 were associated with cancer cell and epithelial cell-specific processes such as regulation of epidermal growth factor-activated receptor activity, positive regulation of angiogenesis, and ERBB2-EGFR signaling pathway (**Figure 10C**). These results suggest that the DORC genes in HCT116 are closely linked to the epithelial cancer cell-specific identity of the cell line. In contrast, DORC genes identified in NK92 were enriched in immune-related processes, including regulation of lymphocyte differentiation and activation, positive regulation of natural killer cell-mediated immunity, and inflammatory response, highlighting the immune cell-specific identity of NK92 cells (**Figure 10D**). To determine whether the peak regions with high gene-peak associations could act as enhancers regulating the expression of DORC genes, we investigated whether these peaks, particularly those linked to cell type-specific gene expression, correspond to super-enhancer regions which are known as the regulatory elements that potentially influence the expression of genes crucial for cell type specification[42]. To compare these DORC regions with the super-enhancer regions of the two cell types, bulk H3K27ac CUT&Tag data for HCT116 and H3K27ac ChIP-seq data for NK92 were used to identify super-enhancers, providing information on chromatin regions activated by the enrichment of active histone markers (**Figures 10E, 10F**). 618 peak regions associated with DORC gene promoters in both cell types overlapped with cell type-specific super-enhancer regions, accounting for approximately 40% of the total peaks (n = 1,505) linked to DORC gene promoters. Notably, when examining the *PLEC* gene of HCT116 DORC and *CCL4* gene of NK92 DORC, it can be observed that snATAC-seq peaks with high associations with DORC gene expression are located near super-enhancer regions (**Figures 11A, 11B**). This suggests that DORC genes are likely critical in defining cell type-specific identity. Furthermore, it indicates that the correlation between cell-type-specific gene expression and chromatin accessibility can be confirmed using SHARE-seq data derived from the same cells.

In addition, it was assessed whether DORC genes linked to a larger number of snATAC-seq peaks have a greater impact on elucidating the identity of the respective cell line and whether they exhibit higher chromatin accessibility compared to other genes. The snATAC-seq peak regions assigned to the previously identified super-enhancer regions were regarded as potential enhancer regions highly correlated with gene expression, suggesting that they may potentially influence gene expression. For the putative enhancer regions, the snATAC-seq signal occupancy of snATAC-seq peak regions assigned to the super-enhancer regions was visualized using the same approach employed to identify the super-enhancer regions. The point where the slope of the signal curve decreased to 1 or below was identified, and putative enhancers located above this cutoff were defined as enhancers with relatively higher accessibility among the putative enhancers. As a result, among the putative enhancer regions containing 56 peak regions in HCT116 and 562 peak regions in NK92, 5 DORC genes in HCT116 and 14 DORC genes in NK92 were associated with regions exhibiting accessibility above the cutoff (**Figures 11C, 11D**). Further investigation of genes linked to enhancers with relatively higher accessibility revealed that DORC genes with a larger number of associated peaks were not necessarily connected to putative enhancers located within the super-enhancer regions. Consequently, it was confirmed that DORC genes with more associated peaks were not always linked to putative enhancers with higher accessibility. Based on the mean DORC scores for each DORC gene across all cells calculated as the sum of normalized scATAC-seq reads aligned at significantly associated DORC peaks, as previously noted, genes associated with peaks of higher accessibility tended to have higher DORC scores compared to genes with a greater number of associated peaks. However, the ranking of genes associated with the most accessible peaks and genes with higher DORC scores (**Figures 12A, 12B**) differed, suggesting that both the number of gene-peak associations and the accessibility of the regions collectively contribute to defining key genes that determine cell identity.

To identify transcription factors (TFs) that could act as putative regulators driving or suppressing the expression of cell type-specific genes as key markers of cell identity, we examined the TF-DORC association based on the enrichment of TF binding motifs in regions associated with each DORC gene and the correlation between the expression of these TFs and their corresponding DORC genes. Ranking TFs based on their mean regulatory scores for all DORC genes revealed that BACH1 and STAT3 emerged as the top activator TFs in HCT116, while OVOL2, ZNF302 and RUNX2 were identified as key

TFs in NK92 cells (**Figures 12C, 12D**). Specifically in NK92 cells, motif searches conducted on the promoter regions of all DORC genes and their associated regulatory regions confirmed that RUNX family TFs were consistently enriched, aligning with the results shown in Figure 12D (**Figure 12E**). RUNX2, in particular, is a known transcriptional regulator essential for NK cell development and maturation[17]. The list of 12 DORC genes (*NEAT1*, *RRM2*, *PIK3AP1*, *GLRX*, *MARS*, *CEACAM21*, *RGS1*, *BMI1*, *KRT80*, *ZNF683*, *GEM*, *LINC00642*) identified as potentially regulated by RUNX2 in NK92 and bulk RNA-seq and RUNX2 ChIP-seq data in NK cells[17] were intersected and found that *PIK3AP1* and *ZNF683* both showed RUNX2 ChIP-seq signals near their genomic loci and were downregulated in RUNX2 knockout NK cells, additionally *PIK3AP1* was upregulated in RUNX2 overexpressing NK cells. These findings suggest that RUNX2 may act as an activator of these genes (**Figures 12F–12H**) in NK cells. The *PIK3AP1* gene is known to activate phosphoinositide 3-kinase (PI3K) in B cells and NK cells, with its role more thoroughly studied in B cells and its potential relevance to NK cell functions, such as target cell recognition and lysis, has been suggested[43]. Meanwhile, *ZNF683* gene plays a critical role in regulating NK cell differentiation, as it is highly upregulated during the differentiation of umbilical cord progenitor cells into NK cells and functions as a transcriptional repressor of interferon-gamma (IFN-γ) production during terminal NK cell differentiation[44]. These findings highlight the possibility that the functions of these two genes, potentially regulated by RUNX2, may play critical roles in defining NK cell identity. In conclusion, this study demonstrates the intricate interplay between chromatin accessibility, gene regulation, and transcription factor activity in defining cell type-specific identities through single-cell multimodal analysis, offering valuable insights into the regulatory mechanisms that underpin cellular identity and function.

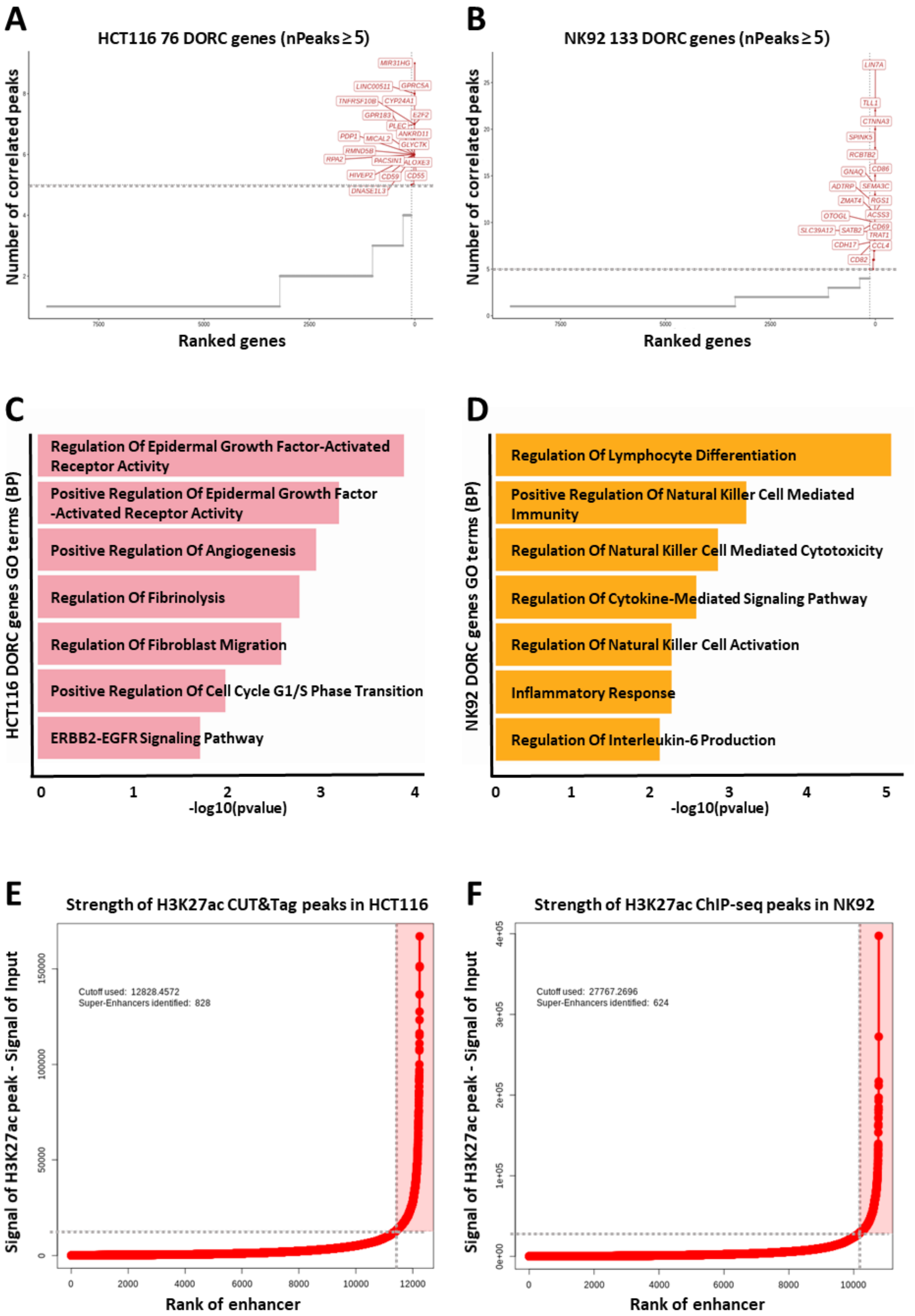
A
HCT116 76 DORC genes (nPeaks≥5)
Number of correlated peaks
Ranked genes
7500
5000
2500
0
MIR31HG
LINC00511
GPRC5A
TNFRSF10B
CYP24A1
GPR183
E2F2
PLEC
ANKRD11
PDP1
MICAL2
GLYCTK
RMND5B
RPA2
PACSIN1
ALOXE3
HIVEP2
CD59
CD55
DNASE1L3
B
NK92 133 DORC genes (nPeaks≥5)
Number of correlated peaks
Ranked genes
25
20
15
10
5
0
7500
5000
2500
0
LIN7A
TLL1
CTNNA3
SPINK5
RCBTB2
GNAQ
CD86
ADTRP
SEMA3C
ZMAT4
RGS1
OTOGL
ACSS3
SLC39A12
SATB2
CD69
TRAT1
CDH17
CCL4
CD82
C
HCT116 DORC genes GO terms (BP)
Regulation Of Epidermal Growth Factor-Activated Receptor Activity
Positive Regulation Of Epidermal Growth Factor -Activated Receptor Activity
Positive Regulation Of Angiogenesis
Regulation Of Fibrinolysis
Regulation Of Fibroblast Migration
Positive Regulation Of Cell Cycle G1/S Phase Transition
ERBB2-EGFR Signaling Pathway
0
1
2
3
4
-log10(pvalue)
D
NK92 DORC genes GO terms (BP)
Regulation Of Lymphocyte Differentiation
Positive Regulation Of Natural Killer Cell Mediated Immunity
Regulation Of Natural Killer Cell Mediated Cytotoxicity
Regulation Of Cytokine-Mediated Signaling Pathway
Regulation Of Natural Killer Cell Activation
Inflammatory Response
Regulation Of Interleukin-6 Production
0
1
2
3
4
5
-log10(pvalue)
E
Strength of H3K27ac CUT&Tag peaks in HCT116
Signal of H3K27ac peak - Signal of Input
Cutoff used: 12828.4572
Super-Enhancers identified: 828
0
50000
100000
150000
0
2000
4000
6000
8000
10000
12000
Rank of enhancer
F
Strength of H3K27ac ChIP-seq peaks in NK92
Signal of H3K27ac peak - Signal of Input
Cutoff used: 27767.2696
Super-Enhancers identified: 624
0e+00
1e+05
2e+05
3e+05
4e+05
0
2000
4000
6000
8000
10000
Rank of enhancer

Figure 10. The DORC genes identified from SHARE-seq data are associated with cell identity determination, exhibiting cell type-specific characteristics through their interactions with regulatory elements. (A) Cell type-specific DORC genes for HCT116 cells (n = 3,937) and the number of snATAC-seq peaks associated with each gene. (B) Cell type-specific DORC genes for NK92 cells (n = 6,036) and the number of snATAC-seq peaks associated with each gene. (C, D) Gene ontology of each cell type-specific DORC gene (E) Strength of H3K27ac CUT&Tag peak signals in HCT116 cells. Super enhancer peaks are colored in red. (F) Strength of H3K27ac ChIP-seq peak signals in NK92 cells. Super enhancer peaks are colored in red.

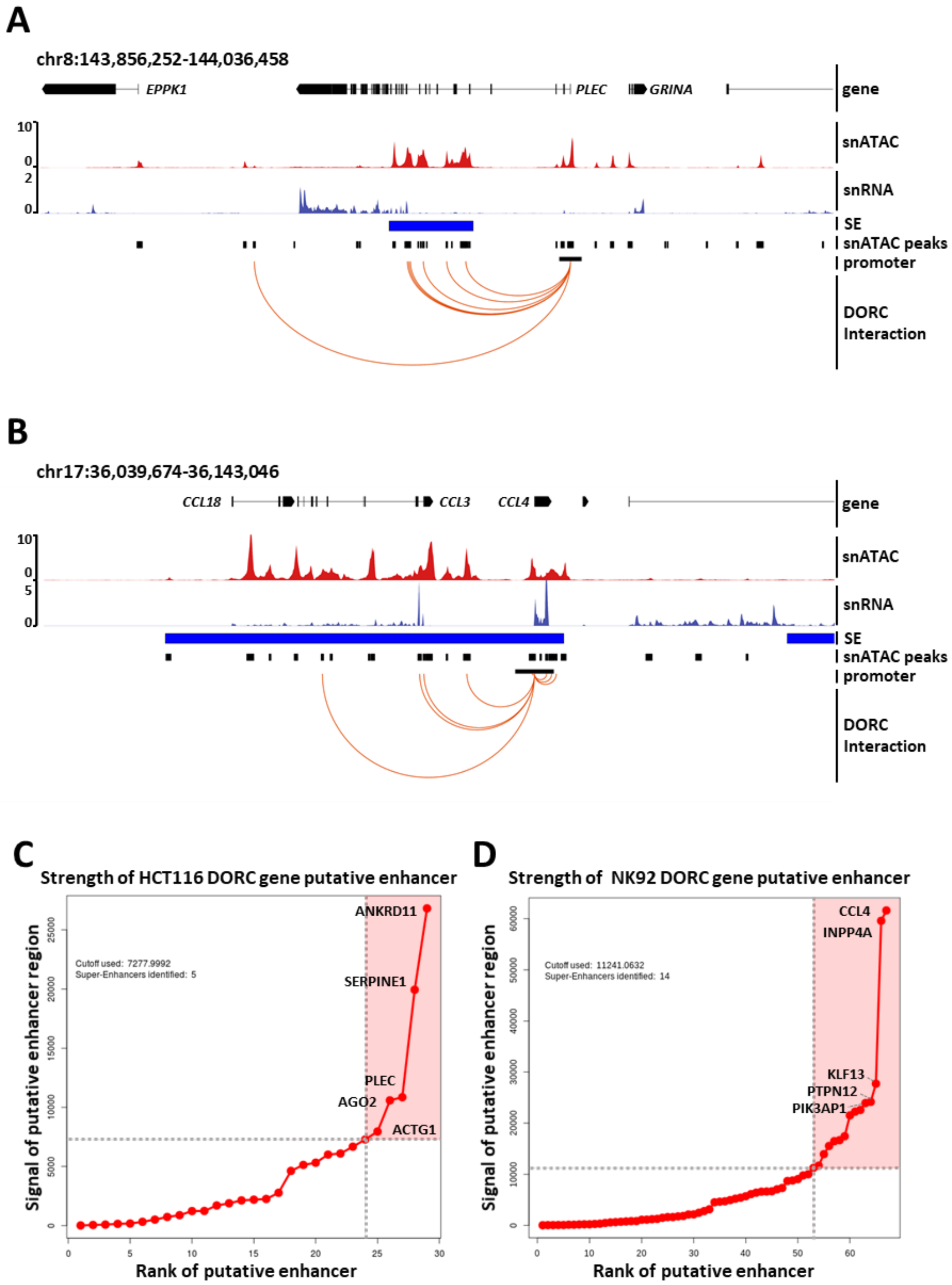
A
chr8:143,856,252-144,036,458
EPPK1
PLEC
GRINA
gene
10
0
snATAC
2
0
snRNA
SE
snATAC peaks
promoter
DORC
Interaction
B
chr17:36,039,674-36,143,046
CCL18
CCL3
CCL4
gene
10
0
snATAC
5
0
snRNA
SE
snATAC peaks
promoter
DORC
Interaction
C
Strength of HCT116 DORC gene putative enhancer
Signal of putative enhancer region
0
5000
10000
15000
20000
25000
Cutoff used: 7277.9992
Super-Enhancers identified: 5
ANKRD11
SERPINE1
PLEC
AGO2
ACTG1
0
5
10
15
20
25
30
Rank of putative enhancer
D
Strength of NK92 DORC gene putative enhancer
Signal of putative enhancer region
0
10000
20000
30000
40000
50000
60000
Cutoff used: 11241.0632
Super-Enhancers identified: 14
CCL4
INPP4A
KLF13
PTPN12
PIK3AP1
0
10
20
30
40
50
60
Rank of putative enhancer

Figure 11. The number of gene-peak associations and the accessibility of the regions collectively contribute to defining key genes that determine cell identity. (A) The interaction between the HCT116 DORC gene *PLEC* and the snATAC-seq peak regions connected to its promoter. (B) The interaction between the NK92 DORC gene *CCL4* and the snATAC-seq peak regions connected to its promoter (C) Strength of putative enhancers of HCT116 DORC genes. Super enhancer-like snATAC-seq peaks (n=5) are colored in red. The DORC genes associated with the top 5 accessible regions are indicated at each point. (D) Strength of putative enhancers of NK92 DORC genes. Super enhancer-like snATAC-seq peaks (n=14) are colored in red. The DORC genes associated with the top 5 accessible regions are indicated at each point.

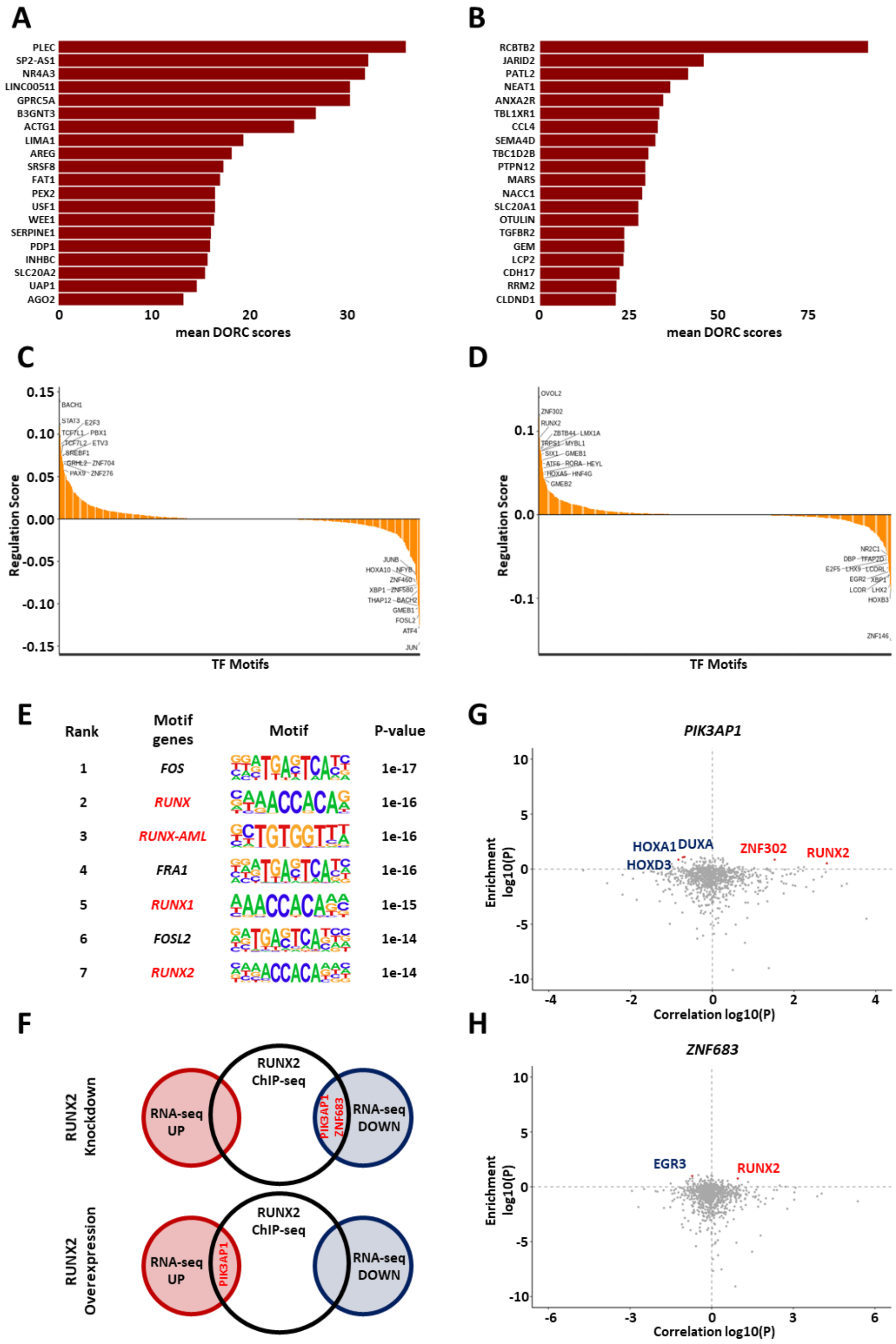
A
PLEC
SP2-AS1
NR4A3
LINC00511
GPRC5A
B3GNT3
ACTG1
LIMA1
AREG
SRSF8
FAT1
PEX2
USF1
WEE1
SERPINE1
PDP1
INHBC
SLC20A2
UAP1
AGO2
0
10
20
30
mean DORC scores
B
RCBTB2
JARID2
PATL2
NEAT1
ANXA2R
TBL1XR1
CCL4
SEMA4D
TBC1D2B
PTPN12
MARS
NACC1
SLC20A1
OTULIN
TGFBR2
GEM
LCP2
CDH17
RRM2
CLDND1
0
25
50
75
mean DORC scores
C
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
Regulation Score
BACH1
STAT3
E2F3
TCF7L1
PBX1
TCF7L2
ETV3
SREBF1
GRHL2
ZNF704
PAX9
ZNF276
JUNB
HOXA10
NFYB
ZNF460
XBP1
ZNF580
THAP12
BACH2
GMEB1
FOSL2
ATF4
JUN
TF Motifs
D
0.1
0.0
-0.1
Regulation Score
OVOL2
ZNF302
RUNX2
ZBTB44
LMX1A
TRPS1
MYBL1
SIX1
GMEB1
ATF6
RORA
HEYL
HOXA5
HNF4G
GMEB2
NR2C1
DBP
TFAP2D
E2F5
LHX9
LCORL
EGR2
XBP1
LCOR
LHX2
HOXB3
ZNF146
TF Motifs
E
Rank
Motif genes
Motif
P-value
1
FOS
1e-17
2
RUNX
1e-16
3
RUNX-AML
1e-16
4
FRA1
1e-16
5
RUNX1
1e-15
6
FOSL2
1e-14
7
RUNX2
1e-14
F
RUNX2 Knockdown
RUNX2 ChIP-seq
RNA-seq UP
RNA-seq DOWN
PIK3AP1
ZNF683
RUNX2 Overexpression
RUNX2 ChIP-seq
RNA-seq UP
RNA-seq DOWN
PIK3AP1
G
PIK3AP1
10
5
0
-5
-10
Enrichment log10(P)
HOXA1
DUXA
HOXD3
ZNF302
RUNX2
-4
-2
0
2
4
Correlation log10(P)
H
ZNF683
10
5
0
-5
-10
Enrichment log10(P)
EGR3
RUNX2
-6
-3
0
3
6
Correlation log10(P)

Figure 12. TF-DORC association based on single cell multimodal data infers putative regulators of cell type-specific gene expression. (A) HCT116 DORC genes that account for the top 20 mean DORC scores across all cells. (B) NK92 DORC genes that account for the top 20 mean DORC scores across all cells. (C) Putative transcription factor drivers of HCT116 are ranked by the overall mean regulation score across all HCT116 DORCs. (D) Putative transcription factor drivers of NK92 are ranked by the overall mean regulation score across all NK92 DORCs. (E) RUNX family motif is enriched in all NK92 DORC gene-associated regions. (F) Genes which can be regulated by RUNX2 are intersected with public RNA-seq and ChIP-seq data from human NK cells in conditions of RUNX2 knockdown and RUNX2 overexpression each. (G, H) The analysis of potential regulatory factors for the *PIK3AP1* gene (G) and *ZNF683* gene (H) suggests that RUNX2 may act as an activator for both genes. The x-axis represents the correlation between the TF and the DORC gene, while the y-axis indicates the degree of enrichment of the corresponding TF motif.

4. Discussion

This study reaffirmed a comprehensive bioinformatics pipeline for analyzing single-cell multiomics data, focusing on the SHARE-seq technique. By applying this pipeline to both public datasets and a mixture of two cell lines datasets (HCT116 and NK92), its robustness and reproducibility were demonstrated. Through the integrative analysis of snRNA-seq and snATAC-seq data obtained from the same cells, we confirmed precise interactions between gene expression and chromatin accessibility, revealing distinct characteristics and regulatory mechanisms between cell types. The analysis of public datasets (human kidney tissue) successfully reproduced transcriptional and epigenomic profiles, confirming the reliability of the pipeline. The clear clustering patterns observed across both snRNA-seq and snATAC-seq modalities closely matched existing study results, showcasing the pipeline's ability to accurately reflect cell-type-specific characteristics. Particularly, the Weighted Nearest Neighbor (WNN) integration approach proved highly effective for enhancing cell-type annotation using multiomics data. In the in-house mixed cell line dataset, distinct transcriptional and chromatin accessibility profiles of HCT116 and NK92 were used to distinguish features between cell states that could not be accurately identified in unimodal analyses. The identification of domains of regulatory chromatin in both HCT116 and NK92 provided key insights into the regulatory mechanisms underpinning cell identity. DORCs discovered in HCT116 were strongly associated with epithelial cancer cell-specific processes, while those identified in NK92 were linked to immune-related functions. These findings demonstrate that SHARE-seq data can elucidate the relationship between regulatory elements and gene expression at a single-cell level. Interestingly, while DORC genes with numerous gene-peak associations were strongly linked to certain super-enhancer regions, indicating that DORCs play a significant role in defining cell identity due to the complex regulatory functions of epigenetic elements. The identification of transcription factors capable of regulating the expression of cell type-specific genes revealed that distinct transcription factors act as key regulators depending on the cell type. This finding reaffirms that the complex interplay between chromatin accessibility, gene regulation, and transcription factor activity collectively contributes to defining cell type-specific identities. However, several challenges remain in single cell multimodal data analysis. First, cells displaying mixed characteristics between cell types were observed during analysis, which could indicate biological intermediate states, necessitating careful consideration during cell type annotation. Particularly, cell embedding based on dimensionality reduction methods can produce highly variable clustering results depending on the number of principal components used. Moreover, marker gene expression analyses relying solely on differentially expressed genes are often heavily influenced by findings from bulk studies, which may hinder precise annotations. Since the choice of marker genes can lead to subjective annotations, it is critical to establish rigorous statistical and biological standards to ensure objective and consistent analyses. Balancing automated annotation methods with manual review will be essential to achieving this goal. Finally, snATAC-seq data alone is insufficient for pinpointing exact enhancer locations. Beyond correlation-based linkage analyses, the incorporation of epigenetic histone modification data, such as H3K27ac and H3K4me1, at a single-cell level is crucial. Integrating such additional data would not only enhance the accuracy of cell type and state definitions but also provide a deeper understanding of the functional roles of regulatory elements.

5. Conclusion

This study validated the bioinformatics pipeline for analyzing single-cell multiomics data and demonstrated its capability to effectively characterize transcriptional and epigenetic features across various cell types and conditions using both public and in-house SHARE-seq datasets. By integrating snRNA-seq and snATAC-seq data from the same cells, the study elucidated interactions between gene expression and chromatin accessibility. The analysis of human kidney tissue data confirmed the high reliability of the pipeline in reproducing transcriptional and epigenomic profiles. Furthermore, the analysis of mixed samples of HCT116 and NK92 cell lines effectively resolved cellular heterogeneity and identified distinct biological characteristics of each cell type. In particular, the identification of domains of regulatory chromatin and their association with super-enhancer regions revealed that chromatin accessibility and gene expression are differentially regulated depending on cell type and play a critical role in determining cellular identity. Furthermore, it was observed that putative enhancers associated with DORC genes containing a large number of connected peaks did not always exhibit the highest accessibility. Instead, the number of gene-peak associations and accessibility collectively influenced the identification of key genes that determine cellular identity. Analysis of TF-DORC associations revealed that distinct transcription factors regulate the expression of cell type-specific genes depending on the cell type. In particular, RUNX2 was suggested to act as an activator TF for two DORC genes associated with NK cell function and differentiation in NK92 cells, reaffirming previously known characteristics of RUNX2. Additionally, the study addressed challenges such as potential subjective bias in cell type annotation and the necessity of incorporating additional data, such as histone modification profiles, to accurately identify enhancer regions. These findings underscore the importance of establishing objective standards for cell annotation and integrating complementary multiomics data to clarify interactions between cis-regulatory elements. In conclusion, this study demonstrates that single-cell multimodal analysis using snRNA-seq and snATAC-seq data can effectively explore cellular heterogeneity and identify key regulatory elements that define cellular identity and function. In particular, the ability of SHARE-seq to integratively analyze transcriptomic and chromatin accessibility features provides valuable insights into the dynamic characteristics and interactions of gene regulation at the single-cell level.

References

1. Nguyen QH, Lukowski SW, Chiu HS, Senabouth A, Bruxner TJC, Christ AN, et al. Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Res 2018;28:1053-66.

2. Calbo J, van Montfort E, Proost N, van Drunen E, Beverloo HB, Meuwissen R, et al. A functional role for tumor cell heterogeneity in a mouse model of small cell lung cancer. Cancer Cell 2011;19:244-56.

3. Tellez-Gabriel M, Ory B, Lamoureux F, Heymann MF, Heymann D. Tumour Heterogeneity: The Key Advantages of Single-Cell Analysis. Int J Mol Sci 2016;17.

4. Zhao Q, Eichten A, Parveen A, Adler C, Huang Y, Wang W, et al. Single-Cell Transcriptome Analyses Reveal Endothelial Cell Heterogeneity in Tumors and Changes following Antiangiogenic Treatment. Cancer Res 2018;78:2370-82.

5. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell 2018;175:598-9.

6. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol 2018;19:621-37.

7. Cramer P. Organization and regulation of gene transcription. Nature 2019;573:45-54.

8. Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet 2019;20:437-55.

9. Schuettengruber B, Bourbon HM, Di Croce L, Cavalli G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell 2017;171:34-57.

10. Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription. Nat Rev Mol Cell Biol 2015;16:178-89.

11. Hegenbarth J-C, Lezzoche G, De Windt LJ, Stoll M. Perspectives on Bulk-Tissue RNA Sequencing and Single-Cell RNA Sequencing for Cardiac Transcriptomics. Frontiers in Molecular Medicine 2022;2.

12. Li X, Wang CY. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci 2021;13:36.

13. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 2023;24:550-72.

14. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell 2020;183:1103-16 e20.

15. Li H, Li D, Ledru N, Xuanyuan Q, Wu H, Asthana A, et al. Transcriptomic, epigenomic, and spatial metabolomic cell profiling redefines regional human kidney anatomy. Cell Metab 2024;36:1105-25 e10.

16. Lee EC, Kim K, Jung WJ, Kim HP. Vorinostat-induced acetylation of RUNX3 reshapes transcriptional profile through long-range enhancer-promoter interactions in natural killer cells. BMB Rep 2023;56:398-403.

17. Wahlen S, Matthijssens F, Van Loocke W, Taveirne S, Kiekens L, Persyn E, et al. The transcription factor RUNX2 drives the generation of human NK cells and promotes tissue residency. Elife 2022;11.

18. Krueger F. Trim Galore! ; 2019.

19. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357-9.

20. Picard toolkit. Broad Institute, GitHub repository: Broad Institute; 2019.

21. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10.

22. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.

23. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.

24. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE

reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766-D73.

25. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884-i90.

26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.

27. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923-30.

28. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 2017;27:491-9.

29. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics 2012;28:2184-5.

30. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.

31. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 2011;27:718-9.

32. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods 2021;18:1333-41.

33. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell 2019;177:1888-902 e21.

34. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573-87 e29.

35. Kartha VK, Duarte FM, Hu Y, Ma S, Chew JG, Lareau CA, et al. Functional inference of gene regulation using single-cell multi-omics. Cell Genom 2022;2.

36. 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.

37. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44:W90-7.

38. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 2013;153:307-19.

39. Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 2013;153:320-34.

40. Lopez-Delisle L, Rabbani L, Wolff J, Bhardwaj V, Backofen R, Gruning B, et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets. Bioinformatics 2021;37:422-3.

41. 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.

42. Pott S, Lieb JD. What are super-enhancers? Nat Genet 2015;47:8-12.

43. Gunesch JT, Angelo LS, Mahapatra S, Deering RP, Kowalko JE, Sleiman P, et al. Genome-wide analyses and functional profiling of human NK cell lines. Mol Immunol 2019;115:64-75.

44. Post M, Cuapio A, Osl M, Lehmann D, Resch U, Davies DM, et al. The Transcription Factor ZNF683/HOBIT Regulates Human NK-Cell Development. Front Immunol 2017;8:535.

Abstract in Korean

암-면역 세포 혼합체에서 단일핵 수준의 염색질 접근성과 유전자 발현 통합 분석

세포 간의 이질성에 관한 이해는 조직의 기능과 질병 진행의 복잡성을 해독하는 데 필수적이다. 단일 세포 다중오믹스 기술인 SHARE-seq은 기존의 여러 세포들의 평균적 프로파일링을 탐색하는 수준의 해석을 넘어, 개별 세포의 후성유전체 및 전사체 환경을 탐색할 수 있는 기회를 제공할 수 있다. 본 연구에서는 SHARE-seq 데이터를 활용하여 후성유전체 이질성을 탐색하기 위한 생물정보학 분석 파이프라인을 검증하고, 이를 자체적으로 생산된 암-면역 세포 혼합체 SHARE-seq 데이터에 적용해 세포 이질성을 확인하였다. 파이프라인의 재현성을 검증하기 위해 공개된 인간 신장 조직의 SHARE-seq 데이터에 적용하였으며, 신장 세포의 전사체 및 후성유전체의 이질성이 성공적으로 재구성되었고, 전사체와 염색질 접근성 프로파일을 바탕으로 기존 연구의 결과와 일치하는 명확한 세포 클러스터를 식별하였다. 또한, 검증된 파이프라인을 활용해 대장암 세포주와 면역 세포주를 혼합한 자체 SHARE-seq 데이터의 두 모달리티를 통합 분석함으로써, 두 개의 세포주를 성공적으로 구분하였다. 나아가, 두 모달리티 간의 연관성이 매우 높은 조절 염색질 영역을 식별하고, 슈퍼 인핸서 영역과의 연관성을 분석함으로써, 염색질 접근성과 유전자 발현이 세포 유형에 따라 다르게 조절되며, 두 요소의 연관성 정도 및 접근성 수준과 같은 여러 요인들이 세포

특이적 유전자의 발현에 함께 영향을 줄 수 있다는 것에 대한 통찰을 얻을 수 있었다. 추가적으로, 세포에 따라 서로 다른 전사 인자의 활성이 존재하여 세포 특이적인 기능을 수행하는 유전자의 발현에 영향을 줌으로써, 염색질 접근성, 유전자 조절, 전사 인자 활성 간의 복잡한 상호작용이 세포 유형 특이적 정체성을 정의하는 데 복합적으로 기여한다는 점을 재확인할 수 있었다. 결론적으로, 본 연구는 단일 세포 다중오믹스 데이터의 통합 분석을 통하여 유전자 발현과 염색질 접근성의 관계를 종합적으로 탐구하였으며, 세포 정체성과 기능을 정의하는 주요 조절 요소를 식별함으로써 더 나아가 암과 같은 질병의 개별 세포 간 이질성 및 조절 환경을 이해하는 데 중요한 통찰을 제공할 수 있을 것으로 기대한다.

핵심되는 말 : 세포이질성, 단일세포 다중오믹스, SHARE-seq, 생물정보학, 통합분석

PUBLICATION LIST

1. Song MJ, Kim M, Seo J, Kwon HW, Yang CH, Joo JS, et al. Role of histone modification in chromatin-mediated transcriptional repression in protozoan parasite Trichomonas vaginalis. BMB Rep 2025.