creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

연세대학교
YONSEI UNIVERSITY

Generation and quality control of single-nucleus multi-omics data in a cancer-immune cell mixture

Heon-Woo Kwon

The Graduate School
Yonsei University
Department of Medical Science

연세대학교
YONSEI UNIVERSITY

Generation and quality control of single-nucleus multi-omics data 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

Heon-Woo Kwon

January 2025

연세대학교
YONSEI UNIVERSITY

This certifies that the Master's Thesis of Heon-Woo Kwon 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

연세대학교
YONSEI UNIVERSITY

ACKNOWLEDGEMENTS

It has been two years since I began my master's program, filled with excitement and dreams for research. Completing this thesis would not have been possible without the support of many people. First, I would like to express my deepest respect and gratitude to my advisor, Professor Hyoung-Pyo Kim. From the beginning to the end of this research, your guidance and warm encouragement provided direction and support. Whenever challenges arose, your advice and words of encouragement helped me grow and discover the joy of research. I also sincerely thank Professor Hyun Seok Kim and Professor Byungjin Hwang for dedicating valuable time and effort to review my thesis. Your meticulous feedback and thoughtful advice significantly improved the quality of my work. To the members of the laboratory who shared this journey, I extend my heartfelt thanks. Special thanks to Dr. Chul Min Yang for establishing the foundational SHARE-seq protocol for this thesis. I am also grateful to MiKyung Kim for offering advice and creating a comfortable lab environment. I deeply appreciate the senior colleagues who generously provided guidance and support throughout my research. Eun-Chong offered the right direction during difficult moments, and Jung-Sik, as a mentor, gave genuine advice and showed continuous interest in my work. Thanks to Hyunsung for patiently answering all my questions and to Sugyung for helping frame the analysis of my research data. My heartfelt thanks go to my fellow cohort, Jieun, who has supported me since the beginning of the program. The analyses

연세대학교
YONSEI UNIVERSITY

conductedallowed my data to be organized into this thesis. I am also grateful to Chang Hoon, who joined the lab with me and provided mutual support as a fellow researcher. Gratitude also extends to Hyejin for offering assistance whenever needed and to Chanyeon for valuable insights on data analysis despite our short time together. Lastly, thank you, Jongcheol. I look forward to seeing the lab's contributions continue to grow. Though I cannot mention everyone here, I wish to extend appreciation to all friends who offered support and encouragement throughout the program. Finally, I express my deepest gratitude to my parents for their boundless love and support, which made this achievement possible. I believe that graduation is not an end but a new beginning. Building upon the knowledge and experience gained during my master's program, I will continue to explore and grow as a researcher.

Winter 2024

With heartfelt gratitude

Heon-Woo Kwon

연세대학교
YONSEI UNIVERSITY

TABLE OF CONTENTS

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

LIST OF FIGURES

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

LIST OF TABLES

연세대학교
YONSEI UNIVERSITY

ABSTRACT

Generation and quality control of single-nucleus multi-omics data in a cancer-immune cell mixture

Single-cell studies have enabled the exploration of cellular heterogeneity, the identification of rare cell types, and the investigation of developmental processes and cell fate. However, single-cell studies focusing on a single modality provide only partial insights into the complex gene regulatory networks within cells. To address these limitations, experimental methods have been developed to simultaneously analyze the genome, epigenome, transcriptome, and proteome within the same cells. Currently, droplet-based methodologies are widely used for multiomics studies, but they are costly and have low throughput. In this study, SHARE-seq (Simultaneous High-throughput ATAC and RNA Expression with Sequencing) was applied to generate and analyze libraries from a mixed sample of the immune cell line NK92 and the colorectal cancer cell line HCT116. SHARE-seq, a combinatorial indexing-based method, offers higher throughput and improved cost efficiency compared to conventional droplet-based multiomics techniques. Using this method, chromatin accessibility and gene expression profiles specific to each cell line were identified at the single-nucleus level within the mixed sample. Furthermore, UMAP analysis revealed distinct clusters corresponding to the NK92 and HCT116 cell lines for each modality. Finally, nuclei with matching barcodes in both snATAC-seq and snRNA-seq clusters were identified. These nuclei represent high-quality samples for further analyses, such as Weighted Nearest Neighbor (WNN) analysis or studies on the

연세대학교
YONSEI UNIVERSITY

functional relationships of regulatory elements controlling gene expression.

This study demonstrates that simultaneous analysis of chromatin accessibility and gene expression at the single-nucleus level in a cancer-immune cell mixture enables the precise distinction between immune and cancer cell lines by leveraging data from two modalities within the same nucleus. Additionally, it highlights the potential to identify high-quality nuclei for future analyses aimed at exploring the functional relationships of regulatory elements governing gene expression. These findings are expected to contribute to the precise identification of cell types and enhance our understanding of cell-cell interactions and gene regulatory networks in complex biological systems, such as the tumor microenvironment (TME). Moving forward, the integration of single-cell multiomics data is anticipated to be widely applied for characterizing and analyzing cell types in tissues composed of diverse cell populations or within specific in vivo environments.

Key words : single nucleus, multimodal sequencing, chromatin accessibility, transcriptome

연세대학교
YONSEI UNIVERSITY

1. INTRODUCTION

1.1. Chromatin dynamics and transcriptional regulation

The central dogma, a fundamental principle in molecular biology, describes the unidirectional flow of genetic information from DNA to RNA to protein[1]. According to this concept, an organism's genome sequence contains all the necessary information to define its state. However, the field of epigenetics emerged to account for biological phenomena that cannot be fully explained within the central dogma's framework. Broadly speaking, epigenetics serves as a bridge between genotype and phenotype, altering gene expression at specific loci or chromosomes without changing the underlying DNA sequence[2]. Among the various epigenetic mechanisms, chromatin accessibility plays a particularly critical role.

Chromatin, composed of DNA and histone proteins, is located in the nucleus of eukaryotic cells and is organized into a tightly packed structure of nucleosomes. Each nucleosome consists of a histone octamer core around which 147 base pairs of DNA are wound[3,4]. Chromatin exists in euchromatic or heterochromatic states, and gene expression is regulated by chromatin accessibility[5]. Accessible chromatin regions across the genome, including enhancers, promoters, insulators, and transcription factor binding sites, collectively control gene expression[6]. In contrast, inaccessible chromatin represents areas where transcription factors cannot bind, leading to minimal transcriptional activity[7]. Currently, ATAC-seq is recognized as a pivotal technique for identifying euchromatin regions[8,9]. This method employs transposase to target accessible chromatin regions. The enzyme cleaves DNA in euchromatin regions, where chromatin structure is open, and

연세대학교
YONSEI UNIVERSITY

inserts adaptors. These DNA fragments are subsequently analyzed using next-generation sequencing (NGS), enabling precise mapping of accessible chromatin locations across the genome. In living organisms, controlling chromatin accessibility is essential for defining cellular identity and function[10]. It regulates the development of stem cells into specific cell types during embryonic development and mediates responses to environmental stimuli and cellular signals[11,12]. ATAC-seq provides valuable insights into how cells regulate their functions and respond to changes. Additionally, it allows researchers to understand how disruptions in these regulatory mechanisms can lead to disease. As a result, ATAC-seq is a crucial tool for deepening our knowledge of chromatin dynamics and its relevance to cellular biology and disease.

Transcription is regulated by various mechanisms, including chromatin accessibility. Among the resulting RNA transcripts, some undergo 5' capping[13], polyadenylation[14,15], and RNA splicing[16,17] to become mature mRNAs ready for protein translation. Regulating mRNA expression enables cells to adapt to external signals and internal needs, ensuring functional diversity and specificity. Currently, mRNA-seq (messenger RNA sequencing) is a widely utilized and powerful technique for transcriptome analysis, providing comprehensive insights into gene expression profiles at the cellular level[18-20]. This method leverages next-generation sequencing (NGS) to quantify and sequence mRNA molecules, enabling the identification of expressed genes, transcript variants, and novel transcripts. By capturing the transcriptome, mRNA-seq provides a representation of the cellular functional state, revealing genes actively being transcribed under specific conditions. Accurate control of mRNA expression is essential for physiological processes like cell differentiation, growth, and stress adaptation. By controlling the timing and specificity of gene expression, cells can execute specialized functions and maintain harmony across tissues and organs.

연세대학교
YONSEI UNIVERSITY

As a robust and versatile tool, mRNA-seq has become a cornerstone in functional genomics, systems biology, and disease modeling. With its ability to provide high-resolution and comprehensive transcriptome data, mRNA-seq continues to advance our understanding of gene regulation and its profound implications for cellular biology and therapeutic development.

1.2. Single-cell multiomics technology

Unlike traditional bulk-level experiments, single-cell RNA sequencing technology has revolutionized molecular biology by enabling unprecedented scale and resolution in transcriptome profiling[20]. With the advent of single-cell transcriptome analysis, efforts have expanded to explore the genome[21], epigenome[22,23] and proteome[24] at the single-cell level, making it possible to analyze these dimensions individually. However, experiments targeting a single modality capture only one aspect of the intricate regulatory elements that control cellular differentiation, function, and signal transduction.

To overcome the limitations of single-cell unimodal sequencing, various experimental approaches have emerged that enable multimodal analysis at single-cell resolution, incorporating diverse modalities such as chromatin accessibility[25-27], histone modifications[28,29], surface protein expression[30,31] and spatial location[32]. Single-cell transcriptomics, the leading single-cell omics approach, is frequently integrated with other omics methods to investigate the link between gene expression and phenotypic heterogeneity without bias[33]. Furthermore, this integrative analysis provides important insights into the interactions between distal regulatory elements, like enhancers, and their target genes, facilitating the study of intercellular communication and regulatory networks at single-cell resolution. Single-cell multiomics has been transformative in developmental

연세대학교
YONSEI UNIVERSITY

biology by enabling lineage tracing and identifying epigenetic changes that drive cell fate decisions. In disease research, it has offered valuable understanding of cancer progression, immune cell heterogeneity in autoimmune diseases, and the molecular mechanisms behind neurodegenerative disorders. The integration of spatially resolved techniques with single-cell multiomics further adds a crucial layer of spatial context, particularly valuable for studying tumor microenvironments and tissue architecture. Progress in multimodal sequencing has enabled the concurrent analysis of various molecular characteristics within individual cells, offering a more comprehensive understanding of how different regulatory layers interact and function together at single-cell resolution.

1.3. SHARE-seq

SHARE-seq (Simultaneous High-throughput ATAC and RNA Expression with Sequencing) is an experimental platform designed to jointly analyze chromatin accessibility and transcriptomic data at the single-cell level[34]. By integrating chromatin accessibility and transcriptomic data generated through SHARE-seq, regulatory elements controlling gene expression can be identified. This joint profiling approach establishes a direct link between transcriptional regulation and its downstream outputs, enabling greater insight into the molecular processes underlying cellular physiology[35]. Moreover, the concurrent measurement of various molecular features in individual cells provides a comprehensive insight into the interactions and functions of regulatory layers within cells. SHARE-seq enables the identification of functional links between regulatory elements that govern gene expression and supports the reconstruction of cellular lineages and differentiation pathways through temporal data analysis. In contrast to traditional droplet-based single-cell separation methods, SHARE-seq employs multiple rounds of

연세대학교
YONSEI UNIVERSITY

hybridization-ligation to simultaneously label mRNA and chromatin fragments originating from the same cell[36]. This process involves three rounds of combinatorial indexing, with 96 unique barcodes ligated to gDNA and cDNA in each round, resulting in a total of 884,736 possible barcode combinations. This high-throughput combinatorial indexing method offers significant advantages over widely used droplet-based single-cell multimodal sequencing, including higher efficiency and cost-effectiveness.

SHARE-seq is not limited to single-cell applications but is also compatible with single-nucleus profiling, enabling its use in tissues where traditional single-cell sequencing is challenging. For instance, adipocytes, which are large and high lipid-rich content, pose significant difficulties for droplet-based single-cell sequencing[37]. Similarly, cardiomyocytes, characterized by their single or dual nucleus configuration and intercalated disc-mediated connectivity, are challenging to sequence at the single-cell level[38]. However, SHARE-seq can be applied at the single-nucleus level, enabling the acquisition of chromatin accessibility and gene expression dataset even from tissues where single-cell sequencing is challenging. In summary, SHARE-seq represents a robust and versatile approach for concurrent profiling of chromatin accessibility and gene expression, enabling high-throughput and cost-effective analyses across diverse sample types, including those from challenging tissues, at both single-cell and single-nucleus resolutions.

연세대학교
YONSEI UNIVERSITY

2. MATERIALS AND METHODS

2.1. Cell line culture

The HCT116 human colorectal cancer cell line (ATCC) was maintained at 37°C in an atmosphere of 5% CO_2 using RPMI 1640 medium (HyClone) supplemented with 10% FBS (HyClone), 100 U/mL penicillin, and 100 μg/mL streptomycin (HyClone). The human NK cell line NK-92 (ATCC) was cultured at 37°C with 5% CO_2 in α-MEM medium (Gibco) supplemented with 20% FBS (HyClone), 55 μM 2-mercaptoethanol (Gibco), 100 U/mL penicillin, 100 μg/mL streptomycin (HyClone), and 100 U/mL IL-2 (Roche).

2.2. SHARE-seq library preparation

SHARE-seq was conducted using methods described in prior studies[34,39,40], with minor modifications introduced by Dr. Chul Min Yang and Dr. Eun-Chong Lee.

2.2.1. Annealing oligo plates

Linker strands and barcode sequences used during the hybridization rounds were prepared in 96-well plates, with each well holding 10 μL of oligos at defined concentrations. In Round 1, the concentrations were 9 μM for the linker and 10 μM for the barcodes; in Round 2, 11 μM for the linker and 12 μM for the barcodes; and in Round 3, 13 μM for the linker and 14 μM for the barcodes. The linker oligos were prepared in STE buffer containing 10 mM Tris-HCl, pH 8.0, 50 mM NaCl, and 1 mM EDTA. Annealing was achieved by heating the plates to 95°C for 2 minutes, followed by gradual cooling to 20°C at a rate of - 1°C per minute. Each round contained 96 barcodes. For the full sequences,

연세대학교
YONSEI UNIVERSITY

refer to Supplementary Table S1 in the SHARE-seq publication[34].

2.2.2. Adaptor annealing

Adaptors were annealed following the manufacturer's instructions. Adaptor A and Adaptor B were annealed separately, differing only in the use of the Read 1 oligo for Adaptor A and the Read 2 oligo for Adaptor B (**Table 1**). For each reaction, 100 μM of the respective Read oligo and 100 μM of the ME oligo (**Table 1**) were prepared in annealing buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl, 1 mM EDTA). Annealing was carried out by heating the plates to 95°C for 2 minutes, then slowly cooling to 20°C at a rate of 1°C per minute. The reaction was finalized with a cooling step at 20°C for 2 minutes. Annealed oligos were stored at -20°C.

Table 1. Oligo sequences for adaptor annealing

Index	Oligo sequence
ME oligo	TCTACACATATTCTCTGTC
Read 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Read 2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

2.2.3. Tn5 transposome assembly

The Tn5 transposome was prepared according to the manufacturer's guidelines. In a PCR tube, 100 μM of the Read 1 adaptor and 100 μM of the Read 2 adaptor were mixed in equal volumes to create the adaptor mixture. An equal volume of the adaptor mixture was then combined with unloaded Tn5 (Diagenode) and gently mixed using a pipette. The assembled Tn5 complex was stored at -20°C.

연세대학교
YONSEI UNIVERSITY

2.2.4. Fixation

Briefly, 1 million cells were collected by centrifugation and resuspended in fresh formaldehyde (Thermo Fisher Scientific) to achieve a final concentration of 0.1%. The sample was incubated at room temperature for 5 minutes with rotation. Quenching solution was added to neutralize the formaldehyde, and the cells were incubated on ice for 5 minutes. The cell pellet was washed twice with 1 mL PBS-2RI (1X PBS, 0.835% BSA, 0.03 U/μL SUPERase RNase Inhibitor, 0.06 U/μL Enzymatics RNase inhibitor).

2.2.5. Nuclei isolation

The cells were lysed in MNIB-2 buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40, 0.1% Tween-20, 0.01% Digitonin) for 3 minutes. This was followed by incubation in MNIB-3 buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.01% Digitonin) on ice for 10 minutes. After the nuclei isolation, the nuclei were washed once with NIB-2RI buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40, 0.03 U/μL SUPERase RNase Inhibitor, 0.06 U/μL Enzymatics RNase inhibitor). The isolated nuclei were counted, and 300,000 nuclei were distributed across four tubes for downstream processing.

2.2.6. Transposition and reverse transcription

The extracted nuclei were suspended in 50 μL of PBS-2RI buffer and transferred to a new tube. Prior to performing tagmentation, a 2x TB buffer (0.066 M Tris-acetate, 0.132 M K-acetate, 0.02 M Mg-acetate, 0.2% NP-40, 32% DMF) was prepared. Next, 150 μL of

연세대학교
YONSEI UNIVERSITY

tagmentation mixture (1x TB buffer, 0.01% Digitonin, 1x Proteinase inhibitor, 1.2 U/μL Enzymatic RNase Inhibitor) was added to the sample, followed by incubation at room temperature for 10 minutes. Subsequently, 4 μL of assembled Tn5 was added, and the sample was aliquoted into PCR tubes at 50 μL per tube. The aliquots were maintained at 37°C with shaking at 500 rpm for 30 minutes. After the transposition step, the samples were washed with NIB-2RI buffer. The washed samples were resuspended in 100 μL of reverse transcription mix (0.3 M Betaine, 571 μM dNTPs, 2.38 μM RT-primer, 4.76 mM DTT, 0.01% Triton X-100, 16.7% PEG 8000, 1x RT buffer, 19.05 U/μL Maxima H Minus Reverse Transcriptase, SUPERase RNase Inhibitor 0.29 U/μL, Enzymatic RNase Inhibitor 0.57 U/μL) and aliquoted into 50 μL portions. The aliquots were heated at 50°C for 10 minutes, followed by three thermal cycles (8°C for 12 seconds, 15°C for 45 seconds, 20°C for 45 seconds, 30°C for 30 seconds, 42°C for 120 seconds, and 50°C for 180 seconds), and then incubated at 50°C for 5 minutes. After completing the reverse transcription, the samples were washed with NIB-2RI buffer.

Table 2. Primer sequence used for reverse transcription

Index	Primer sequence
RT_ primer	/5Phos/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG[10-bp UMI]/iBiodT/TTTTTTTTTTTTTTVN

2.2.7. Hybridization and ligation

The samples resuspended in NIB-2RI were mixed with Hybridization buffer (1.67x T4 ligase buffer, 0.17% NP-40, 0.084 U/μL SUPERase RNase Inhibitor, 0.533 U/μL Enzymatic RNase Inhibitor) and dispensed into each well of the Round 1 plate at 40 μL per well. The plate was incubated at 450 rpm and 24°C for 30 minutes. Subsequently, 10 μL

연세대학교
YONSEI UNIVERSITY

of Blocking Oligo 1 was dispensed into each well, followed by incubation of the plate at 450 rpm and 24°C for 30 minutes. After incubation, all samples were transferred to a reservoir and redistributed into the Round 2 plate at 55 µL per well, followed by incubation at 450 rpm and 24°C for 30 minutes. Next, 10 µL of Blocking Oligo 2 was dispensed into each well, followed by incubating the plate for 30 minutes under the same conditions. Finally, the samples were transferred back to a reservoir and distributed into the Round 3 plate at 65 µL per well, followed by a final incubation at 450 rpm and 24°C for 30 minutes. The barcoded samples were washed with NIB-2RI buffer and resuspended in 80 µL of NIB-2RI buffer. The samples were subsequently suspended in 320 µL of ligation mixture (1.25x T4 Ligase buffer, 0.125% NP-40, 25 U/µL T4 DNA ligase, 0.0625 U/µL SUPERase RNase Inhibitor, 0.4 U/µL Enzymatic RNase Inhibitor) and incubated at 24°C with shaking at 450 rpm for 30 minutes. After completing the ligation step, the samples were washed with NIB-2RI buffer and aliquoted into 50 µL sublibraries at 20,000 nuclei per aliquot. The 20,000-nuclei sublibraries were stored in a -80°C deep freezer.

Table 3. Oligo sequences for combinatorial indexing

Index	Oligo sequence
R1 barcodes	/5Phos/ CGCGCTGCATACTTG[8-bp Barcode1]CCCATGATCGTCCGA
R1 linker	CCGAGCCCACGAGACTCGGACGATCATGGG
R2 barcodes	/5Phos/CATCGGCGTACGACT[8-bp Barcode2]ATCCACGTGCTTGAG
R2 linker	CAAGTATGCAGCGCGCTCAAGCACGTGGAT
R3 barcodes	CAAGCAGAAGACGGCATACGAGAT[8-bp Barcode3]GTGGCCGATGTTTCG
R3 linker	AGTCGTACGCCGATGCGAAACATCGGCCAC
R1 blocking	CCCATGATCGTCCGAGTCTCGTGGGCTCGG
R2 blocking	ATCCACGTGCTTGAGCGCGCTGCATACTTG

연세대학교
YONSEI UNIVERSITY

R3 blocking	GTGGCCGATGTTTCGCATCGGCGTACGACT

2.2.8. Reverse crosslinking and affinity pull-down

For the 20,000 nuclei stored in 50 µL, 1x reverse crosslinking buffer (prepared by diluting 2x reverse crosslinking buffer: 100 mM Tris-HCl, pH 8.0, 100 mM NaCl, 0.004% SDS), 0.2 µg/µL Proteinase K (New England Biolabs), and 0.4 U/µL SUPERase RNase Inhibitor (Thermo Fisher Scientific) were added. The mixture was incubated at 450 rpm at 55°C for 1 hour. After incubation, 5 µL of 100 mM PMSF was added to inactivate Proteinase K, followed by incubation at room temperature for 10 minutes. For affinity pull-down preparation, Dynabeads MyOne Streptavidin T1 (Invitrogen) were washed twice with 1x B&W-T buffer (5 mM Tris-HCl, pH 8.0, 1 M NaCl, 0.5 mM EDTA, and 0.05% Tween 20) and once with 1x B&W-T buffer supplemented with 0.8 U/µL SUPERase RNase Inhibitor (Thermo Fisher Scientific). Add the prepared beads to the sample after Proteinase K inactivation, and incubate with rotation at 10 rpm at room temperature for 60 minutes.

2.2.9. snATAC-seq library preparation

The transposed DNA in the supernatant was purified using the QIAGEN MinElute PCR Purification Kit and eluted with 22 µL of QIAGEN Elution Buffer. The fragments were amplified in a 50 µL PCR reaction containing 1x NEBNext buffer (New England Biolabs), 0.5 µM library-specific Ad1 primer, and 0.5 µM P7 primer. The PCR reaction was performed under the following conditions: 72°C for 5 minutes, 98°C for 30 seconds, followed by 5 cycles of 98°C for 10 seconds, 63°C for 30 seconds, and 72°C for 1 minute. A quantitative PCR was conducted to estimate the number of additional cycles needed for

연세대학교
YONSEI UNIVERSITY

library amplification. This was done using 1 µL of the pre-PCR sample in a total reaction volume of 10 µL. The amplified library, following additional PCR, was purified using the QIAGEN MinElute PCR Purification Kit. The final libraries underwent size selection with 0.9X AMPure XP beads (Beckman Coulter) and were sequenced on the Illumina NovaSeq X platform with the following read specifications: Read 1 – 50 bp, Read 2 – 50 bp, Index 1 – 99 bp, and Index 2 – 8 bp.

2.2.10. cDNA library preparation

After the supernatant (snATAC-seq library) was removed, the beads were washed three times with 1x B&W-T buffer containing 0.2 U/µL SUPERase RNase Inhibitor (Thermo Fisher Scientific) and once with STE buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl, and 1 mM EDTA). The washed beads were suspended in 50 µL of template switch mix containing 1 mM dNTPs, 1 M Betaine, 10% PEG 8000, 1x Maxima RT buffer, 2% Ficoll PM-400, 4 U/µL NxGen RNase Inhibitor, 2.5 µM TSO, and 10.12 U/µL Maxima H Minus Reverse Transcriptase. The mixture was incubated with rotation at 10 rpm at room temperature for 30 minutes, followed by shaking at 300 rpm at 42°C for 90 minutes. After the TSO reaction, 100 µL of distilled water was added, and the beads were washed with STE buffer. The cDNA was amplified in a 50 µL PCR reaction containing template DNA, 1x KAPA HiFi HotStart ReadyMix, 0.4 µM RNA primer, 0.4 µM P7 primer. The PCR reaction was performed under the following conditions: 95°C for 3 minutes, followed by 5 cycles of 98°C for 20 seconds, 65°C for 45 seconds, and 72°C for 3 minutes; then an additional 5 cycles of 98°C for 20 seconds, 67°C for 20 seconds, and 72°C for 5 minutes; and a final extension at 72°C for 5 minutes. To determine the number of additional cycles required for library amplification, a quantitative PCR was performed using a 1 µL aliquot

연세대학교
YONSEI UNIVERSITY

of the PCR product in a total reaction volume of 10 μL containing 1x EvaGreen (Biotium). Based on the qPCR results, the remaining samples were amplified with additional PCR cycles under the conditions of 95°C for 3 minutes, followed by the additional cycles of 98°C for 20 seconds, 67°C for 20 seconds, 72°C for 1 minute, and a final extension at 72°C for 5 minutes.

Table 4. Primer sequences used for cDNA library preparation

Index	Primer sequence
TSO	AAGCAGTGGTATCAACGCAGAGTGAATrGrG+G
RNA primer	AAGCAGTGGTATCAACGCAGAGT

2.2.11. Tagmentation and snRNA-seq library preparation

A tagmentation mixture was prepared with 1x TD buffer (prepared by diluting 2x TD buffer consisting of 20 mM Tris-HCl, pH 7.6, 10 mM $MgCl_2$, 20% dimethylformamide with distilled water), 50 ng of cDNA, and distilled water to a final volume of 100 μL. Subsequently, 10 μL of a 1:80 diluted ME-A adaptor-loaded Tn5 transposase was added, and the reaction was incubated at 55°C with shaking at 300 rpm for 5 minutes. The tagmented samples were purified using the QIAGEN MinElute PCR Clean-Up Kit and eluted in 22 μL of QIAGEN Elution Buffer. The purified samples were amplified through a 50 μL PCR reaction with the following thermal cycling conditions: 72°C for 5 minutes and 98°C for 30 seconds, followed by 7 cycles of 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 1 minute, with a final extension at 72°C for 5 minutes. The final libraries were size-selected using 0.7x AMPure XP beads (Beckman Coulter) and sequenced on the Illumina NovaSeq X platform with the following specifications: Read 1 – 50 bp, Read 2 –

연세대학교
YONSEI UNIVERSITY

50 bp, Index 1 – 99 bp, and Index 2 – 8 bp.

2.3. SHARE-seq library quality control

2.3.1. Polymerase chain reaction and electrophoresis for library size distribution analysis

The obtained library was quantified, and 100 pg of the SHARE-seq library was used for PCR. The PCR conditions were as follows: an initial denaturation at 95°C for 2 minutes, followed by 20 cycles consisting of denaturation at 95°C for 20 seconds, annealing at 63°C for the snATAC-seq library or 67°C for the snRNA-seq library for 30 seconds, and extension at 72°C for 1 minute. A final extension step was performed at 72°C for 5 minutes. The resulting libraries were evaluated via agarose gel electrophoresis to verify their size distribution.

Table 5. Primer sequences for library size distribution analysis

Index	Primer sequence
Illumina P5	AATGATACGGCGACCACCGAGATCTACAC
Illumina P7	CAAGCAGAAGACGGCATACGAGAT
Read 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Read 2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
Index 1	CTGTCTCTTATACACATCTCCGAGCCCACGAGAC

2.3.2. TA cloning and DNA elution for library sequence confirmation

DNA amplified with Illumina P5 and Illumina P7 primers (**Table 5**) was purified using the Expin™ CleanUp SV (GeneAll). The purified DNA insert was cloned into a TA vector (Enzynomics), and transformation was performed in DH5α competent cells

연세대학교
YONSEI UNIVERSITY

(Enzynomics) according to the manufacturer's instructions. Libraries from the resulting colonies were purified using the Exprep™ plasmid SV (GeneAll) and sequenced by Sanger sequencing with the universal primers listed in **Table 6**.

Table 6. Sequencing primer for reading DNA sequence after cloning

Index	Primer sequence
M13R-pUC	CAGGAAACAGCTATGAC

2.4. Bioinformatic analysis

Most of the bioinformatic analyses of the omics data generated in this dissertation were conducted by Jieun Seo. The HCT116_CMV bulk mRNA-seq and ATAC-seq datasets used in this dissertation were generated by Dr. Bobae Yang, while the NK92 bulk mRNA-seq and ATAC-seq datasets were generated by Dr. Eun-Chong Lee.

2.4.1. mRNA-seq data processing

Paired-end sequencing reads were processed using Trim Galore[41] (v0.6.10) with the command-line option trim_galore --paired for adapter trimming. The trimmed reads were aligned to the Human hg38 genome assembly using STAR (v2.5.2b) with the parameters --chimSegmentMin 20 --twopassMode Basic --quantMode TranscriptomeSAM. Gene expression quantification was performed with RSEM[42] (v1.3.1) using the options --paired-end --estimated-rspd. Differentially expressed protein-coding genes were identified using the DESeq2 R package[43] (v1.44.0), applying a $\log_2$ fold-change cutoff of 2 and an adjusted p-value threshold of 0.01. Strand-specific reads were extracted with SAMtools (v1.19.2)[44] and normalized to generate strand-specific mRNA-seq genome tracks using the

연세대학교
YONSEI UNIVERSITY

bamCoverage function from deepTools (v3.5.5)[45] with the normalization method –normalizeUsing CPM.

2.4.2. ATAC-seq data processing

Paired-end sequencing reads were trimmed using Trim Galore with the same parameters applied in mRNA-seq preprocessing. The trimmed reads were aligned to the Human hg38 reference genome using Bowtie2 (v2.5.3) with the settings --end-to-end --very-sensitive --maxins 2000. Reads with low mapping quality, duplicates, and mitochondrial origin were identified and filtered out using SAMtools and Picard Tools (v2.14.1). Nucleosome-free regions were selected, and adaptor insertion sites induced by Tn5 transposase were adjusted with the alignmentSieve function from deepTools, using the command-line options --maxFragmentLength 140 –ATACshift. Nucleosome-free reads were normalized with deepTools in the same way as for ChIP-seq data to produce genome-wide ATAC-seq signal tracks. Peak calling for ATAC-seq data was performed with MACS2 (v2.2.9.1)[46] without incorporating input control data. Differentially accessible regions were identified using the DESeq2 R package, based on read counts from each sample and customized size factors that accounted for the proportion of nucleosome-free reads between samples, with thresholds set at an adjusted p-value of 0.01 and a $\log_2$ fold-change of 2.

2.4.3. SHARE-seq data pre-processing

Pre-processing of SHARE-seq data (.fastq.gz) was performed using previously described scripts (available at https://github.com/masai1116/SHARE-seq-alignmentV2/)[34]. Gene annotation and sequence files (Genome Reference Consortium Human Build 38 patch

연세대학교
YONSEI UNIVERSITY

release 13; GRCh38.p13) from the GENCODE[47] website were used. 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[48] (v0.23.4). For snRNA-seq data, due to the characteristic presence of polyA tails in mRNA, the read2 sequence was excluded, and only the read1 FASTQ file was aligned to the reference genome using STAR[49] (v.2.5.2b). The number of reads mapped to genomic regions was quantified using FeatureCount[50] (v2.0.6), and unique UMI-based read grouping was performed using UMI-tools[51] (v1.1.5) to obtain unique reads by removing duplicated reads. For snATAC-seq (SHARE-ATAC) data, alignment was performed using bowtie2[52] (v2.5.3). Reads that were unmapped, not primary aligned, or aligned to chrM and chrY were removed. Barcodes with fewer than 50 reads were filtered out. The read distribution was checked using RseQC[53] (v5.0.2). This process resulted in a count matrix (.h5 file) representing gene expression and a fragment profile (.bed file) for each cell. To process the .h5 files for generating count matrices, the scanpy.read_10x_h5 function from Scanpy[54] (v1.9.8) was used.

2.4.4. snRNA-seq data processing

All snRNA-seq analysis were executed on Scanpy. Cells with fewer than 1,000 or more than 6,500 genes detected, as well as cells with fewer than 1,000 reads or more than 20,000 reads, were removed from the gene count matrix. Genes present in fewer than 50 cells were also excluded. Cells with more than 30% mitochondrial reads were removed, and doublet detection was performed using the scanpy.external.pp.scrublet function. The expected doublet rate was set to 0.06, and the number of neighbors was set to 30. Cells with doublet scores exceeding 0.2 were annotated as suspected doublets and excluded from

연세대학교
YONSEI UNIVERSITY

analysis. The data were subsequently normalized and log-transformed. Highly variable genes (7,788 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'). A neighborhood graph was computed using the scanpy.pp.neighbors function with the number of neighbors set to 15 (metric='cosine'). The neighborhood graph was embedded into two dimensions using the scanpy.tl.umap function, with the minimum effective distance between embedded points set to 0.5. Leiden clustering was performed using the scanpy.tl.leiden function. For single-cell cluster annotation, a set of marker genes was compiled. Each marker gene was qualitatively visualized in UMAP space to confirm its spatial distribution.

2.4.5. snATAC-seq data processing

Chromatin analysis was conducted using the CreateChromatinAssay function from Signac to generate a chromatin assay from the count matrix, followed by conversion into a Seurat object using the CreateSeuratObject function from Seurat[55] (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 hg38 genome blacklist were calculated using the NucleosomeSignal, TSS Enrichment, FRiP, and FractionCountsInRegion functions, respectively. For quality control, cells with 2,000 to 50,000 peaks, a nucleosome signal value below 2.5, a TSS enrichment score above 4, a FRiP value greater than 0.1, and a blacklist overlap ratio below 0.05 were retained for

연세대학교
YONSEI UNIVERSITY

downstream analysis. The data was then normalized using the TF-IDF (term frequency-inverse document frequency) method implemented in the RunTFIDF function. Singular value decomposition (SVD) was performed on the TF-IDF matrix for linear dimensionality reduction using the RunSVD function. Graph-based clustering, non-linear dimensionality reduction, and UMAP visualization were performed using the FindNeighbors, FindClusters, and RunUMAP functions, respectively, with parameters dims = 2:30, min.dist = 0.5, and n.neighbors = 30. Notably, cells with relatively low FRiP values were carefully excluded during the analysis to avoid artifacts associated with low-FRiP clusters. 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). Peak calling was repeated for each cluster, resulting in the identification of a total of 184,399 features. All snATAC-seq analyses described above were based on the previously constructed peak-by-cell matrix.

연세대학교
YONSEI UNIVERSITY

3. RESULTS

3.1. SHARE-seq workflow to concurrently profile chromatin accessibility and gene expression in a cancer-immune cell mixture.

SHARE-seq (Simultaneous High-throughput ATAC and RNA Expression with Sequencing) is an innovative multiomics platform that allows for concurrent analysis of chromatin accessibility and gene expression at single-cell resolution[34]. In SHARE-seq, cells are first fixed, and their nuclei are isolated. Subsequently, the Tn5 transposase tags regions of open chromatin in the DNA. mRNA is reverse-transcribed using poly(T) primers that include unique molecular identifiers (UMIs) and biotin tags. The transposed DNA and poly(T) cDNA undergo three rounds of hybridization-ligation with 8-bp barcodes in 96-well plates. This process creates 884,736 unique barcode combinations, each of which labels a single nucleus. Reverse crosslinking releases both transposed DNA and poly(T) cDNA, ensuring that each carries the same barcode corresponding to the same cell. The poly(T) cDNA, tagged with biotin, is isolated using streptavidin beads, while the transposed DNA remains in the supernatant. These paired profiles are subsequently identified by matching the well-specific barcode combinations, ensuring that the chromatin accessibility and transcriptomic data are correctly linked for each individual cell (**Figure 1A**).

In this study, the nuclei of the NK92 cell line (immune cells) and the HCT116 cell line (colon cancer cells) were mixed. SHARE-seq was performed on the mixed nuclei to determine whether the two cell lines could be distinguished at the single-nucleus level. Each cell line was fixed, and nuclei were isolated from 1 million cells. Prior to tagmentation, 150,000 nuclei from each cell line were combined, resulting in a total of 300,000 nuclei for

연세대학교
YONSEI UNIVERSITY

tagmentation. Finally, sublibraries were constructed from 20,000 barcoded samples. Using SHARE-seq, chromatin accessibility and gene expression were simultaneously profiled at single-nucleus resolution in the cancer-immune cell mixture, demonstrating that immune and cancer cell lines could be effectively distinguished.

연세대학교
YONSEI UNIVERSITY

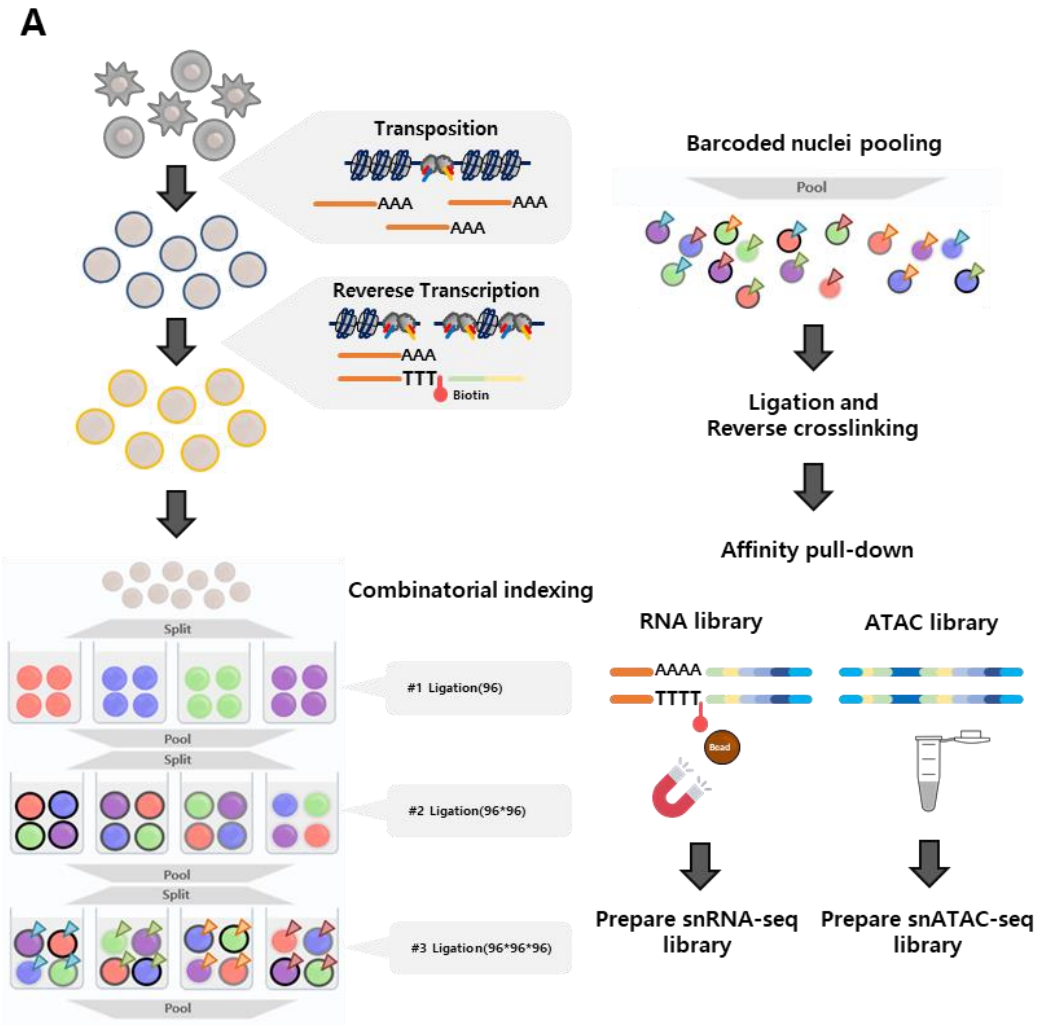


Figure 1. Workflow of SHARE-seq. (A) Schematic representation of SHARE-seq workflow.

연세대학교
YONSEI UNIVERSITY

3.2. SHARE-seq library quality control via polymerase chain reaction and TA cloning.

To separate individual nuclei through combinatorial indexing, the library structure generated by SHARE-seq is more complex compared to conventional single-cell sequencing libraries (**Figure 2A, 2B**). These constructs include well-specific barcodes (BC1, BC2, and BC3), linker sequences, and molecular identifiers (for the snRNA-seq library) to enable the accurate identification of paired profiles. By employing three barcodes, SHARE-seq integrates both chromatin accessibility and gene expression data from individual cells at single-nucleus resolution.

To confirm the accurate assembly of the SHARE-seq library structure, the barcode construct length and sequence composition generated during snRNA-seq and snATAC-seq were analyzed using electrophoresis (**Figure 2C, 2D**). PCR was conducted using primers targeting the sequencing primers of the SHARE-seq library (Read 1, Read 2, and Index 1) and linker sequences that bind to the Illumina flow cell (Illumina P5 and Illumina P7). For samples amplified with combinations of primers Illumina P5 and Illumina P7, Read1 and Illumina P7, or Read2 and Illumina P5, smeared bands were observed during electrophoresis, indicating the presence of inserts in the amplified DNA. In contrast, samples amplified with Index 1 and Illumina P7, targeting only the barcode region, showed a single, distinct band. This result confirms that the three rounds of barcoding were successfully completed and that the library structures within the SHARE-seq libraries were correctly assembled.

To ensure no sequence alterations occurred in the SHARE-seq library, the constructed libraries were amplified with Illumina P5 and P7 primers and subjected to TA cloning. After extracting the transformed libraries on a per-colony basis, sanger sequencing was

연세대학교
YONSEI UNIVERSITY

conducted to evaluate the structural integrity of the libraries and their concordance with the reference sequences. The analysis demonstrated strong concordance between the snRNA-seq and snATAC-seq library sequences and their respective references, confirming the correct library structure (**Figure 2E, 2F**). This result validates the reliability of the barcoding and library construction processes in SHARE-seq.

연세대학교
YONSEI UNIVERSITY

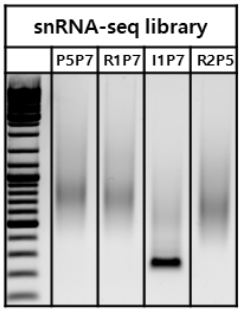
A

snRNA-seq library structure

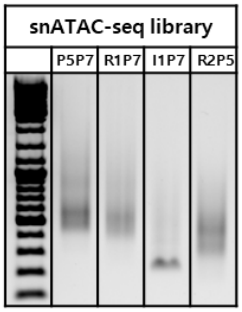
B

snATAC-seq library structure

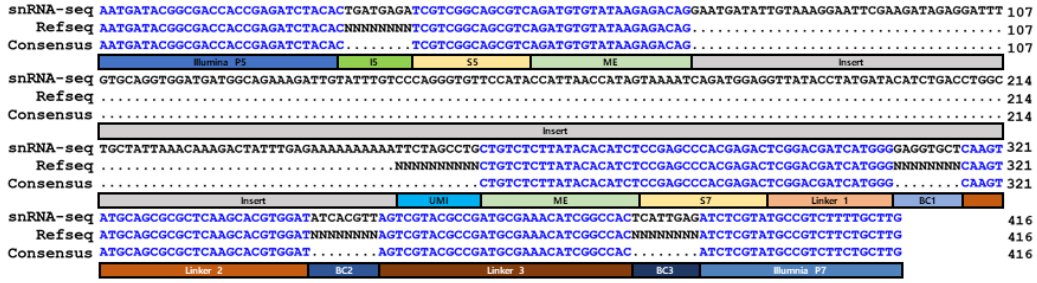
C



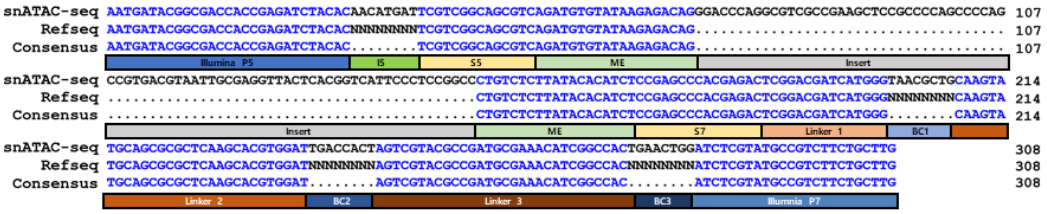
D



E



F



연세대학교
YONSEI UNIVERSITY

Figure 2. Quality control of SHARE-seq libraries using gel electrophoresis and TA cloning. (A) Schematic overview of the snRNA-seq library construction (B) Schematic overview of the snATAC-seq library construction (This schematic overview was adapted and modified with reference to the original overview designed by Dr. Chul Min Yang). (C) The distribution of DNA fragments in snRNA-seq libraries was visualized by gel electrophoresis. (D) The distribution of DNA fragments in snATAC-seq libraries was visualized by gel electrophoresis. (E) Comparison of actual sequences from snRNA-seq libraries obtained via TA cloning with reference sequences. (F) Comparison of actual sequences from snATAC-seq libraries obtained via TA cloning with reference sequences.

연세대학교
YONSEI UNIVERSITY

3.3. Validation of SHARE-seq library quality using bulk sequencing.

To assess the reliability of snRNA-seq and snATAC-seq data generated by SHARE-seq from a cancer-immune cell mixture, a portion of the library underwent bulk sequencing to evaluate its overall quality. Bulk sequencing provides information about the inserts within each library but does not capture barcode information, offering insights into all cell types within the sample. Initially, TapeStation analysis was performed to examine the insert size distribution of the SHARE-seq library. The generated SHARE-seq libraries conformed well to established criteria[40], indicating their high quality (**Figure 3A, 3B**). Additionally, quality control at the bulk level confirmed the high quality of SHARE-seq libraries (**Table 7, 8**).

Further quality control assessments revealed that snRNA-seq datasets typically include a significant proportion of unspliced RNA, resulting in a large number of reads from intronic regions. Bulk sequencing of the snRNA-seq library from the cancer-immune cell mixture revealed an intron rate of approximately 40% (**Figure 3C**), which is significantly higher than that typically observed in mRNA-seq experiments[56]. Additionally, bulk sequencing of the snATAC-seq library showed distinct insert size distributions, with a clear separation of nucleosome-free regions (NFR) at ≤147 bp and mononucleosomes (**Figure 3D**). These findings collectively confirm that the SHARE-seq library was properly constructed at the bulk level and meets the quality requirements for subsequent single-nucleus sequencing.

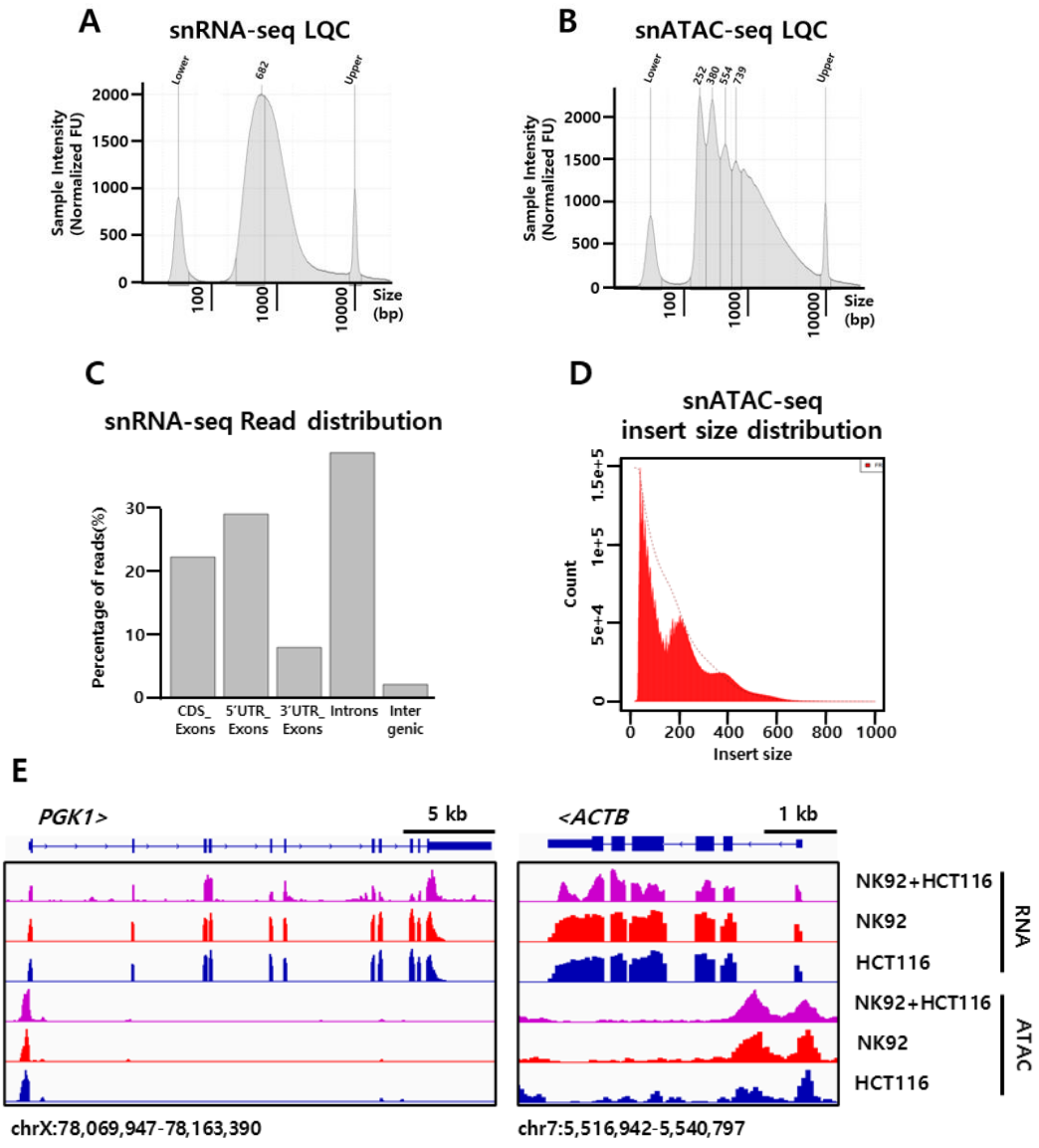
To further validate the SHARE-seq library, RNA and ATAC signals for housekeeping genes were compared to previously generated bulk mRNA-seq and bulk ATAC-seq data from NK92 and HCT116 cells. Genome browser tracks of housekeeping genes, including *PGK1* (Phosphoglycerate Kinase 1) and *ACTB* (Actin Beta), were evaluated through IGV

연세대학교
YONSEI UNIVERSITY

visualization (**Figure 3E**). In the NK92 and HCT116 cell lines, a comparison of signals from SHARE-seq bulk sequencing and bulk-level mRNA-seq and ATAC-seq at housekeeping genes revealed a high level of consistency between the two experiments.

These results confirm the reliability and accuracy of the SHARE-seq library, demonstrating its consistency with bulk sequencing. Furthermore, the SHARE-seq library was confirmed to possess the quality required for single-nucleus sequencing.

연세대학교
YONSEI UNIVERSITY



연세대학교
YONSEI UNIVERSITY

Figure 3. Quality control of SHARE-seq libraries via bulk sequencing. (A) Fragment distribution of snRNA-seq libraries as visualized by TapeStation HS D5000 electropherogram. (B) Fragment distribution of snATAC-seq libraries as visualized by TapeStation HS D5000 electropherogram. (C) Distribution of reads in the bulk-sequenced snRNA-seq library inserts was analyzed. (D) Insert size distribution of the bulk-sequenced snATAC-seq library. (E) Genome tracks of bulk-sequenced SHARE-seq libraries and bulk mRNA-seq and ATAC-seq from NK92 and HCT116_CMV cell lines were visualized in IGV focusing on housekeeping gene regions.

연세대학교
YONSEI UNIVERSITY

Table 7. Quality of snRNA-seq libraries validated by bulk sequencing

Uniquely mapping reads (%)	Percent of reads mapped to multiple loci	Duplication rate (%)
98.13	21.38	42.17

Table 8. Quality of snATAC-seq libraries validated by bulk sequencing

Mapping rate(%)	Duplication rate(%)	FRiP (%)	FRiB (%)	Number Peaks
98.13	21.38	62.17	0.53	72,195

연세대학교
YONSEI UNIVERSITY

3.4. Bulk mRNA-seq and ATAC-seq provide integrated insights into the cancer-immune cell mixture.

Bulk mRNA-seq and ATAC-seq aggregate signals from all cell types in a sample, providing a global view of transcriptional profiles and chromatin dynamics. However, they lack the resolution necessary to differentiate signals from specific cell types. Differential gene expression (**Figure 4A**) and differential accessible region analyses (**Figure 4B**) were conducted using bulk mRNA-seq and ATAC-seq datasets generated for the NK92 and HCT116_CMV cell lines in our laboratory. This allowed for the identification of genes uniquely expressed and active in each cell line, which were subsequently compared with data obtained from bulk-sequenced SHARE-seq libraries.

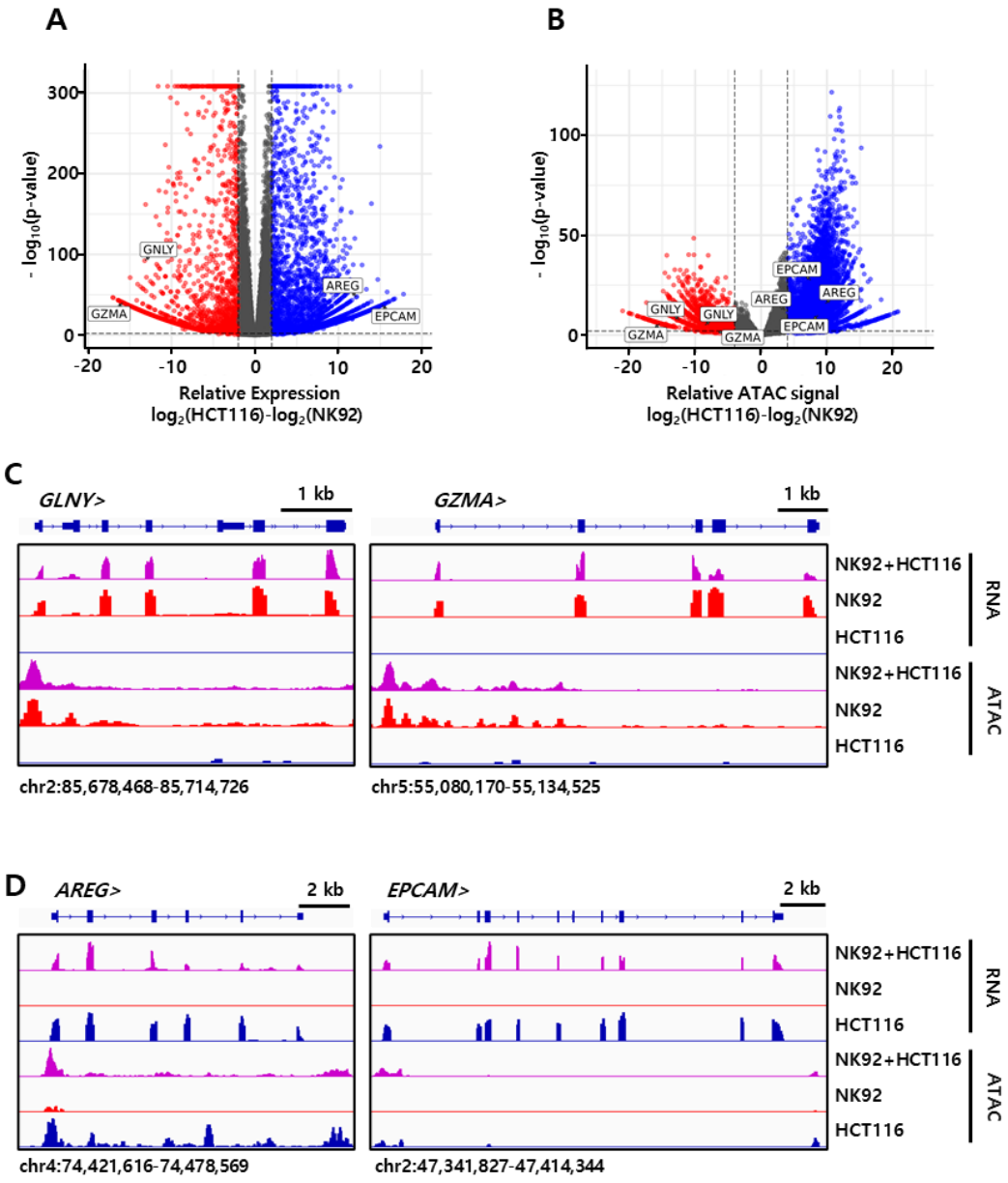
For NK92 cells, *GZMA* and *GNLY* were selected as markers due to their high expression levels, the presence of ATAC-seq signals in accessible promoter regions to NK92 cell, and the absence of similar characteristics in HCT116 cells. Granzyme A (*GZMA*) is abundantly expressed in NK92 cells, inducing caspase-independent cell death by targeting the SET complex to cause DNA damage[57]. Granulysin (*GNLY*) is also highly expressed in NK92 cells, inducing lysis or apoptosis in target cells, tumor cells, or cells infected by intracellular pathogens[58]. These attributes made *GZMA* and *GNLY* ideal signature genes for NK92 cells. Similarly, for HCT116 cells, genes *AREG* and *EPCAM* were selected based on analogous criteria. Amphiregulin (*AREG*) is highly expressed in HCT116 cells, mediating EGFR signaling to drive key oncogenic traits[59]. *EPCAM* is a transmembrane glycoprotein associated with cell-cell adhesion, playing a critical role in tumorigenesis and metastasis[60]. These genes were chosen as HCT116 signature genes due to their specific characteristics.

연세대학교
YONSEI UNIVERSITY

To verify that bulk sequencing represents integrated data from multiple cell types within a sample, IGV genome tracks were utilized to compare signals from bulk mRNA-seq and ATAC-seq datasets with those from bulk-sequenced SHARE-seq libraries. For NK92 signature genes, signals were exclusively observed in NK92 bulk mRNA-seq and ATAC-seq datasets, with no detectable RNA signals or ATAC peaks in the HCT116 datasets. In the SHARE-seq libraries generated from a cancer-immune cell mixture, both RNA signals and ATAC peaks for these signature genes were detected, confirming successful signal integration from both cell types **(Figure 4C**). Similarly, analysis of HCT116 signature genes, such as *AREG* and *EPCAM*, revealed that RNA signals and ATAC peaks were only present in HCT116 bulk datasets, with no corresponding signals in NK92 datasets. However, in the SHARE-seq libraries from a cancer-immune cell mixture, both RNA signals and ATAC peaks for these HCT116 signature genes were detected, further demonstrating effective signal integration from both cell populations **(Figure 4D**).

These results confirm that NK92 and HCT116 cell lines were effectively mixed during SHARE-seq preparation, with the libraries capturing the transcriptome and chromatin dynamics of the mixed sample. In summary, when bulk sequencing was performed on the SHARE-seq library generated from a cancer-immune cell mixture, it provided an overview of the average transcriptome and chromatin dynamics across the mixed cell types in the sample. However, it failed to resolve the transcriptome and chromatin dynamics specific to each individual cell type. These findings highlight the inherent limitations of bulk sequencing in analyzing heterogeneous cell populations.

연세대학교
YONSEI UNIVERSITY



연세대학교
YONSEI UNIVERSITY

Figure 4. Combined cell type signals in bulk sequencing data from the SHARE-seq library. (A) Volcano plot showing differential gene expression analysis of NK92 and HCT116_CMV cell lines using bulk mRNA-seq data. (B) Volcano plot showing differential analysis of accessible regions between NK92 and HCT116_CMV cell lines using bulk ATAC-seq data. (C) Genome tracks of bulk-sequenced SHARE-seq libraries and bulk mRNA-seq and ATAC-seq from NK92 and HCT116_CMV cell lines were visualized in IGV focusing on NK92 cell line signature genes. (D) Genome tracks of bulk-sequenced SHARE-seq libraries and bulk mRNA-seq and ATAC-seq from NK92 and HCT116_CMV cell lines were visualized in IGV focusing on HCT116 cell line signature genes.

연세대학교
YONSEI UNIVERSITY

3.5. Assessment of quality control metrics for snRNA-seq libraries.

To assess the quality and reliability of the snRNA-seq library, comprehensive quality control was performed (**Figure 5A, 5B**). Scanpy[54] was used to conduct quality control and downstream analysis of snRNA-seq data. Key metrics analyzed included the number of genes detected per nucleus (n_genes_by_counts), total read counts (total_counts), mitochondrial read percentage (pct_counts_mt), and doublet rate (doublet_score). SHARE-seq relies on combinatorial indexing to distinguish individual nuclei, which results in the generation of barcodes that do not correspond to actual nuclei. In this experiment, 736,448 barcodes were recognized during the analysis. Therefore, it is crucial to apply quality control measures to filter out non-nuclear barcodes and focus on actual nuclei.

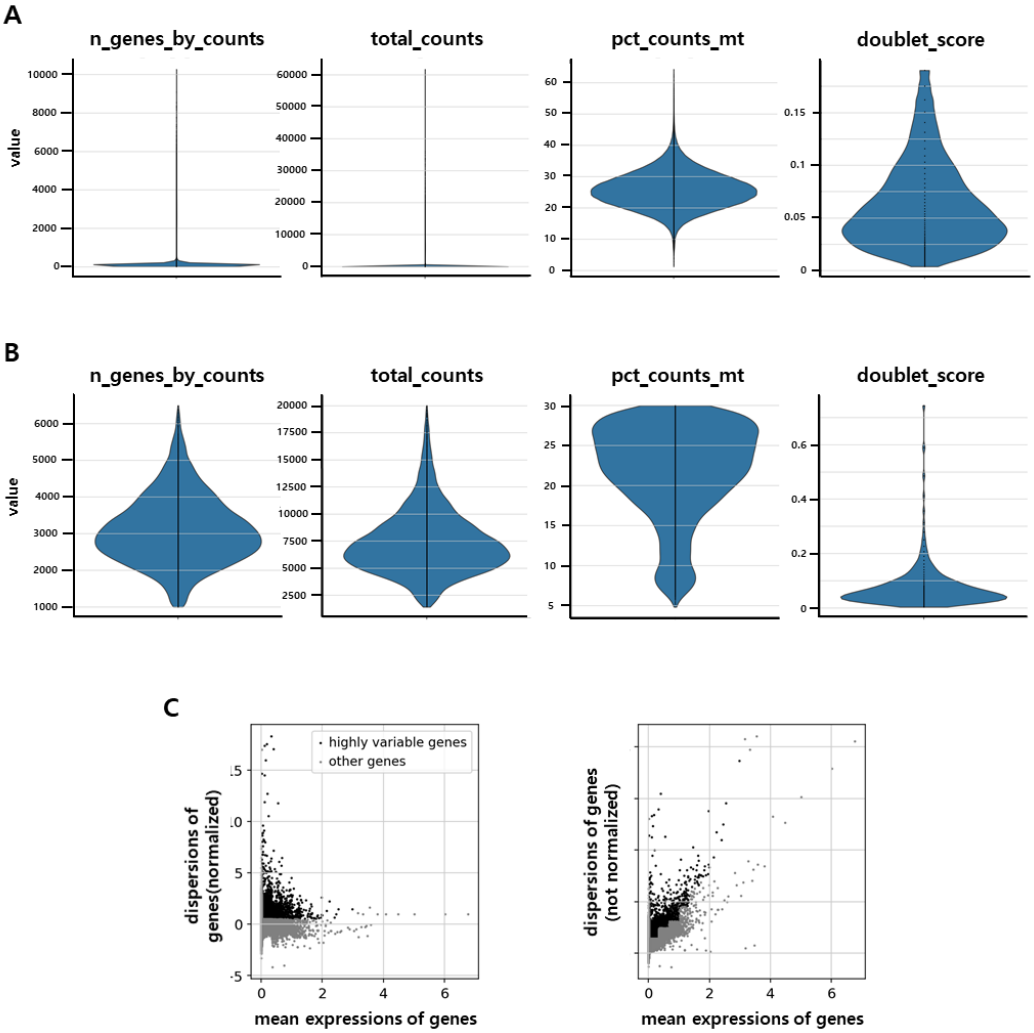
The cutoff for the number of detected genes per nucleus (n_genes_by_counts) was set between 1,000 and 6,500. Nuclei with a gene count outside this range were considered abnormal nuclei and excluded from further analysis. For the number of reads per nucleus (total_counts), a cutoff range of 1,000 to 20,000 was established. Nuclei with total read counts below this range were presumed to represent non-nuclear barcodes rather than actual nuclei and were excluded. The mitochondrial read percentage (pct_counts_mt) was limited to less than 30%, as a high mitochondrial RNA proportion could indicate stressed or compromised nuclei. Finally, doublet scores were restricted to below 0.2%. Doublets, an artifact where two or more nuclei are labeled with a single barcode, were excluded from the analysis. A key advantage of SHARE-seq is its use of combinatorial indexing to label individual nuclei, leading to a much lower doublet rate compared to traditional droplet-based methods[34]. For example, droplet-based methods have an estimated doublet rate ranging from 1% to 10%, depending on the number of cells and the platform used[61]. Conversely, the snRNA-seq data produced in this study exhibited a remarkably low doublet

연세대학교
YONSEI UNIVERSITY

rate, highlighting the method's robustness.

By applying these quality control measures, we successfully removed non-nuclear barcodes and retained 14,610 nuclei for subsequent analysis. This process was essential for eliminating barcode noise inherent to SHARE-seq and ensuring that only experimental nuclei were analyzed. Consequently, this dataset provides a reliable foundation for downstream analyses aimed at understanding transcriptomes at the single-nucleus level. For cells that passed quality control, we performed a feature selection process to identify highly variable genes expressed in nuclei. Specifically, genes were chosen according to the following criteria: normalized mean expression values between 0.0125 and 3, and dispersion values of at least 0.5. Through this process, 7,788 highly variable genes were identified and subsequently utilized for downstream analyses (**Figure 5C**).

연세대학교
YONSEI UNIVERSITY



연세대학교
YONSEI UNIVERSITY

Figure 5. snRNA-seq quality control metrics. (A) Quality control of snRNA-seq data analyzed using Scanpy was performed by filtering nuclei with 1,000–6,500 detected genes (n_genes_by_counts), total counts of 1,000–20,000, mitochondrial gene percentage (pct_counts_mt) below 30%, and a doublet score below 0.2, retaining only 14,610 nuclei for downstream analysis. (B) Violin plots depicting the distribution of quality control metrics for retained nuclei after filtering, including the number of detected genes (n_genes_by_counts), total counts, percentage of mitochondrial gene expression (pct_counts_mt), and doublet scores. (C) Identification of highly variable genes based on their mean expression (x-axis) and dispersion (y-axis) using cutoffs of $0.0125 \leq$ mean expression ≤ 3 and $0.5 \leq$ dispersion, resulting in the selection of 7,788 highly variable genes for downstream analysis.

연세대학교
YONSEI UNIVERSITY

3.6. Assessment of quality control metrics for snATAC-seq libraries.

To evaluate the quality and reliability of the snATAC-seq library, comprehensive quality control was conducted on the snATAC-seq data (**Figure 6A, 6B**). Signac[62] was used to conduct quality control and downstream analysis of snATAC-seq data. Key metrics analyzed included the number of detected peaks per nucleus (nCount_peaks), transcription start site enrichment (TSS.enrichment), blacklist ratio (blacklist_fraction), nucleosome signal (nucleosome_signal), and the proportion of reads within peaks (Pct_reads_in_peaks). SHARE-seq relies on combinatorial indexing to distinguish individual nuclei, which results in the generation of barcodes that do not correspond to actual nuclei. In this experiment, 884,378 barcodes were recognized during the analysis. Therefore, it is crucial to apply quality control measures to filter out non-nuclear barcodes and focus on actual nuclei for subsequent analyses.

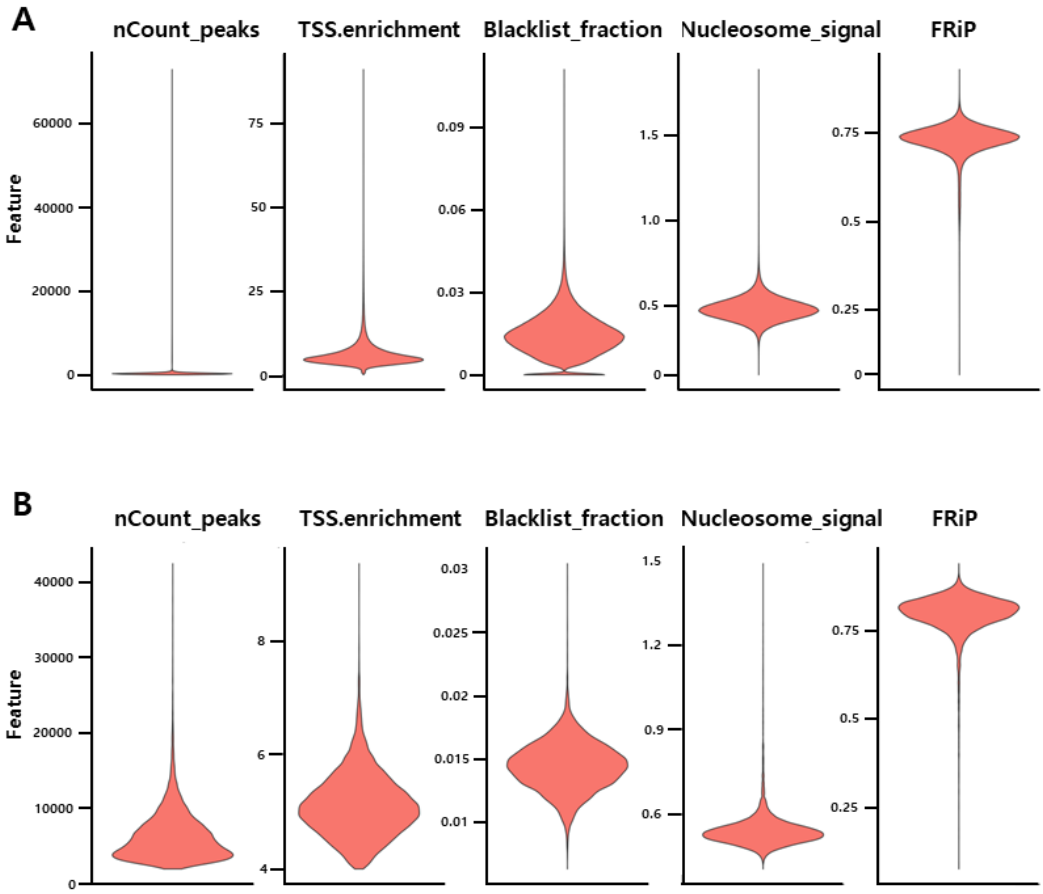
The cutoff for the number of detected peaks per nucleus (nCount_peaks) was set between 2,000 and 50,000. Nuclei with a peak count outside this range were excluded, as they likely represented poorly barcoded nuclei. The TSS enrichment value, which quantifies the signal-to-noise ratio at transcription start sites, was required to exceed a cutoff of 4. Nuclei meeting this threshold were considered to have undergone successful ATAC-seq and were included in further analyses. The blacklist ratio (blacklist_fraction), reflecting the proportion of reads mapping to artifact-prone genomic regions, was limited to below 0.05 to exclude spurious signals. Additionally, the nucleosome signal, which assesses whether tagmentation predominantly occurred in nucleosome-free regions (NFR), was set at less than 2.5. This metric, calculated as the ratio of mononucleosome reads to NFR reads, ensured that nuclei with high-quality ATAC-seq data targeting euchromatic regions were selected. Finally, the fraction of reads in peak (FRiP) was set to a minimum of 0.1,

연세대학교
YONSEI UNIVERSITY

indicating a well-constructed library with minimal background noise and robust peak detection.

By applying these quality control measures, we successfully removed barcode artifacts and retained 17,833 nuclei for subsequent analysis. This process was critical for mitigating barcode noise inherent to SHARE-seq and ensuring that only experimental nuclei were analyzed. Consequently, this dataset provides a robust foundation for downstream analyses aimed at understanding chromatin dynamics at the single-nucleus level. This quality control process underscores the reliability of the dataset for advanced investigations into chromatin accessibility and its regulatory implications.

연세대학교
YONSEI UNIVERSITY



연세대학교
YONSEI UNIVERSITY

Figure 6. snATAC-seq quality control metrics. (A) Quality control of snATAC-seq data was performed by filtering nuclei with 2,000–50,000 peaks detected (nCount_peaks), transcription start site enrichment (TSS.enrichment) greater than 4, blacklist fraction below 0.05, nucleosome signal below 2.5, and fraction of reads in peaks (FRiP) above 0.1, retaining only 17,833 nuclei for downstream analysis. (B) Violin plots showing the distribution of quality control metrics for retained nuclei after filtering, including the number of detected peaks (nCount_peaks), TSS enrichment, blacklist fraction, nucleosome signal and fraction of reads in peak (FRiP).

연세대학교
YONSEI UNIVERSITY

3.7. Defining individual cell types within a sample through gene expression and chromatin accessibility using SHARE-seq.

From the initial 20,000 nuclei processed using SHARE-seq, we applied quality control criteria to identify 14,610 nuclei from the snRNA-seq data and 17,833 nuclei from the snATAC-seq data. These high-quality nuclei were used for downstream analyses, including visualizing clusters of nuclei with similar characteristics in the SHARE-seq data using UMAP[63]. **Figure 7A** shows UMAP clustering based on the snRNA-seq data, revealing three distinct clusters: Cluster 0, Cluster 1, and Cluster 2. Similarly, **Figure 7B** shows the UMAP clustering from the snATAC-seq data, identifying two distinct clusters: Cluster 0 and Cluster 1. Given that the SHARE-seq experiment involved a mixture of cancer and immune cells, we hypothesized that the clusters from each modality correspond to NK92 and HCT116 nuclei. To prioritize snRNA-seq clusters, the expression levels of genes differentially expressed across clusters were examined (**Figure 7C**). This analysis identified the top 200 genes with the highest statistical significance for each cluster, which were used in subsequent analyses. Cell typing was performed on the snRNA-seq data using Panglao DB[64] and the ARCHS4 Cell Lines database[65]. Additionally, functional characteristics and pathway enrichment for each cluster were analyzed using the Elsevier Pathway Collection.

Cluster 0 exhibited significant enrichment in pathways related to cancer cell motility, invasion, and survival, such as "Integrins in Cancer Cell Motility, Invasion, and Survival" and "Proteins with Altered Expression in Cancer Metastasis", as determined by the Elsevier Pathway Collection. Cell typing using the ARCHS4 Cell Lines database identified HCT116 and CPAC1 cell lines as representative of this cluster. These findings suggest that Cluster

연세대학교
YONSEI UNIVERSITY

0 predominantly represents HCT116 cells (**Figure 8A**).

Cluster 1 showed strong enrichment for immune-related pathways, particularly those associated with NK cells, such as "Natural Killer Cell Activation through ITAM-Containing Receptors" and "Natural Killer Cell Precursor - Natural Killer Cell Surface Expression Markers." Cell type analysis with Panglao DB identified NK cell-related populations, such as natural killer cells. These results indicate that Cluster 1 represents NK92 cells (**Figure 8B**).

Cluster 2 was enriched in pathways related to translation, including "Translation and rRNA Translation and Processing". Cell typing using the ARCHS4 Cell Lines database identified HCT116 and SKOV3 cell lines. Cluster 2 was characterized by high ribosomal RNA expression, as the top 200 uniquely expressed genes included numerous ribosomal RNA genes. These findings suggest that Cluster 2 exhibits the expression of some genes characteristic of HCT116. However, the majority of ribosomal RNA genes predominantly represent the features of Cluster 2. This observation indicates the presence of nuclei contaminated with ribosomal RNA in a subset of HCT116 cells (**Figure 8C**).

In contrast, cell type classification and pathway analysis of the top 200 uniquely active genes in the snATAC-seq data failed to identify specific cell types. The challenge in performing cell typing using gene activity stems from the inherent nature of ATAC-seq, which measures chromatin accessibility rather than direct gene expression. Consequently, it infers gene activity indirectly, leading to lower accuracy in ranking the expressed genes. Next, we visualized the expression profiles of signature genes for NK92 and HCT116 cell lines on UMAP plots for both the snRNA-seq and snATAC-seq datasets and validated these findings using dot plots. Signature genes for NK92 and HCT116 were selected based on differentially expressed genes (DEG) and gene expression levels derived from bulk

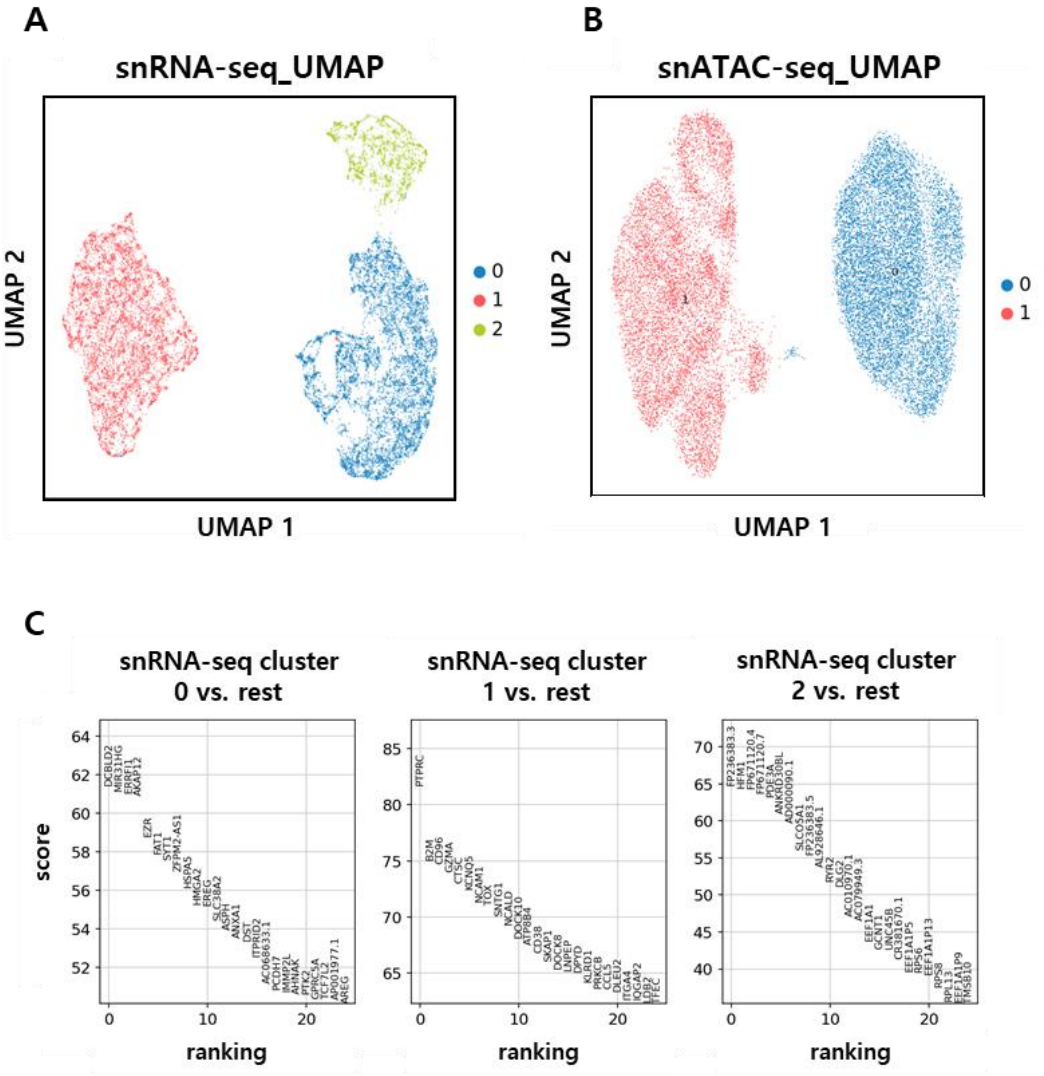
연세대학교
YONSEI UNIVERSITY

mRNA-seq data. In the snRNA-seq data, NK92 signature genes showed higher expression in Cluster 1, with dot plots demonstrating that the average expression levels and the number of cells expressing these signature genes were higher compared to Clusters 0 and 2 (**Figure 9A, 9B**). Similarly, HCT116 signature genes showed higher expression in Clusters 0 and 2, with dot plots confirming that both the average expression levels and the number of cells expressing these signature genes were higher compared to Cluster 1 (**Figure 10A, 10B**). These observations led to the classification of Cluster 1 as NK92 and Cluster 0 as HCT116 in the snRNA-seq data. In Cluster 2, although the expression of HCT116 signature genes was observed to be high, the preceding analysis revealed that ribosomal RNA constituted the majority of the cluster's representative genes. Therefore, Cluster 2 was classified as an rRNA-enriched HCT116 cluster.

We performed a similar analysis on the snATAC-seq data using the previously identified signature genes. NK92 signature genes exhibited higher activity in Cluster 1, with dot plots indicating that cells with high signature gene activity were more abundant in Cluster 1 compared to Cluster 0 (**Figure 11A, 11B**). Conversely, HCT116 signature genes showed higher activity in Cluster 0, with dot plots confirming that cells with high gene activity were more prevalent in Cluster 0 than in Cluster 1 (**Figure 12A, 12B**). These results led us to classify Cluster 0 as representing the HCT116 cell line and Cluster 1 as representing the NK92 cell line in the snATAC-seq data.

In summary, the analysis classified NK92 and HCT116 clusters in both snRNA-seq and snATAC-seq data, clearly mapping cell types based on gene expression and chromatin accessibility.

연세대학교
YONSEI UNIVERSITY

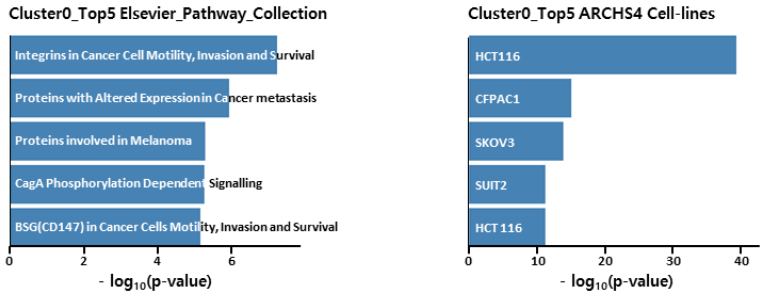


연세대학교
YONSEI UNIVERSITY

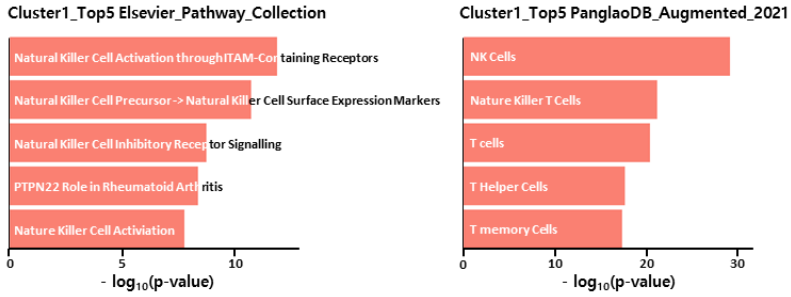
Figure 7. UMAP visualization of SHARE-seq data. (A) SHARE-seq UMAP plot of single nuclei from a mixed HCT116 and NK92 cell line sample, with UMAP coordinates based on snRNA-seq data. (B) SHARE-seq UMAP plot of single nuclei from a mixed HCT116 and NK92 cell line sample, with UMAP coordinates derived from snATAC-seq data. (C) Top 25 signature genes for each of the three clusters from snRNA-seq analysis, ranked by statistical significance (Wilcoxon test, $-\log_{10}$(p-value)).

연세대학교
YONSEI UNIVERSITY

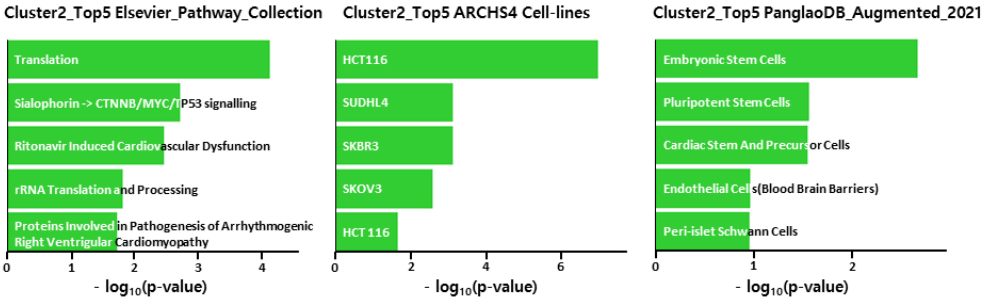
A



B



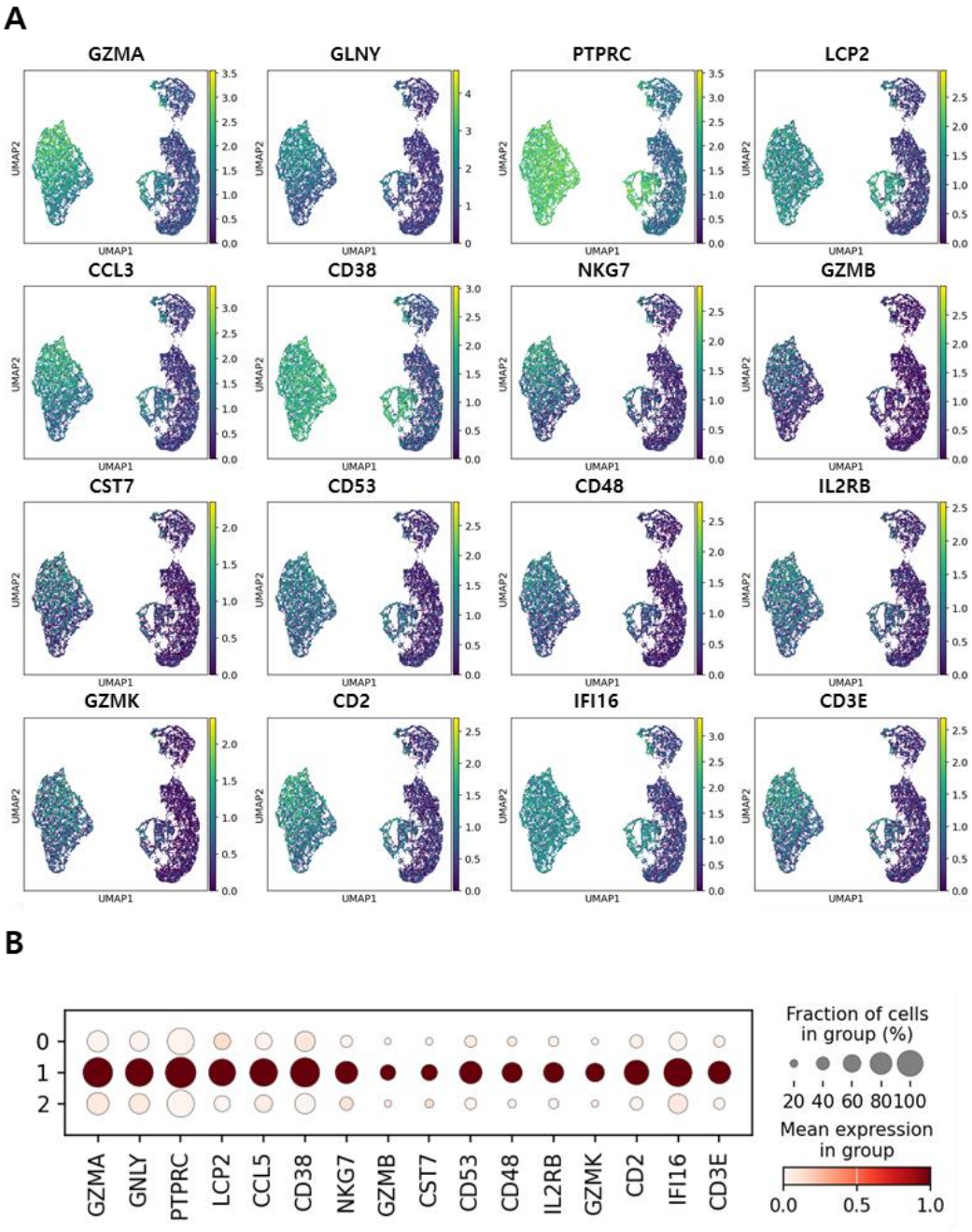
C



연세대학교
YONSEI UNIVERSITY

Figure 8. Cell typing of snRNA-seq clusters. (A) Elsevier Pathway Collection and ARCHS4 Cell lines analysis for 200 genes specifically expressed in Cluster 0 from snRNA-seq, visualizing the top 5 pathways ranked by - $\log_{10}$(P-value) and the top 5 related cell lines ranked by -$\log_{10}$(P-value). (B) Elsevier Pathway Collection and PanglaoDB_Augmented_2021 analysis for 200 genes specifically expressed in Cluster 1 from snRNA-seq, visualizing the top 5 pathways ranked by - $\log_{10}$(P-value) and the top 5 related cell lines ranked by -$\log_{10}$(P-value). (C) Elsevier Pathway Collection, ARCHS4 Cell lines, PanglaoDB_Augmented_2021 analysis for 200 genes specifically expressed in Cluster 2 from snRNA-seq, visualizing the top 5 pathways ranked by -log10(P-value) and the top 5 related cell lines ranked by - $\log_{10}$(P-value).

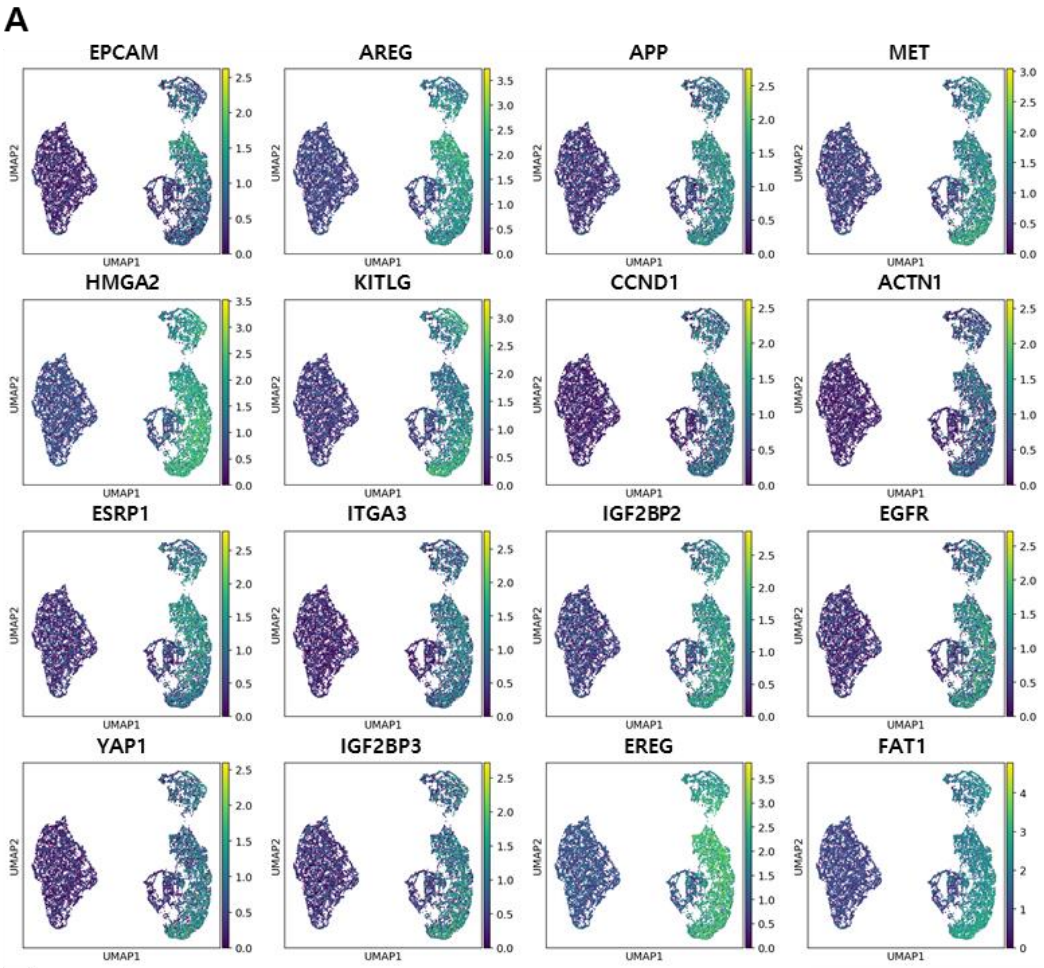
연세대학교
YONSEI UNIVERSITY

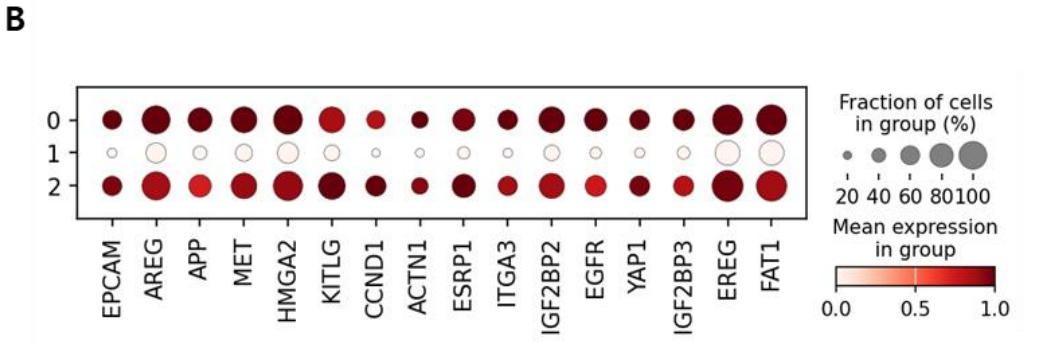


연세대학교
YONSEI UNIVERSITY

Figure 9. Expression of NK92 cell line signature genes in snRNA-seq clusters. (A) UMAP projection showing the expression patterns of 16 NK92 signature genes across snRNA-seq clusters. The NK92 signature genes were identified through bulk mRNA-seq analysis of NK92 and HCT116_CMV, selecting genes that are exclusively expressed in the NK92 cell line and are highly expressed with functional relevance to NK92. (B) Dot plot showing the fraction of cells (dot size) and mean expression levels (color intensity) of 16 NK92 signature genes across snRNA-seq clusters.

연세대학교
YONSEI UNIVERSITY

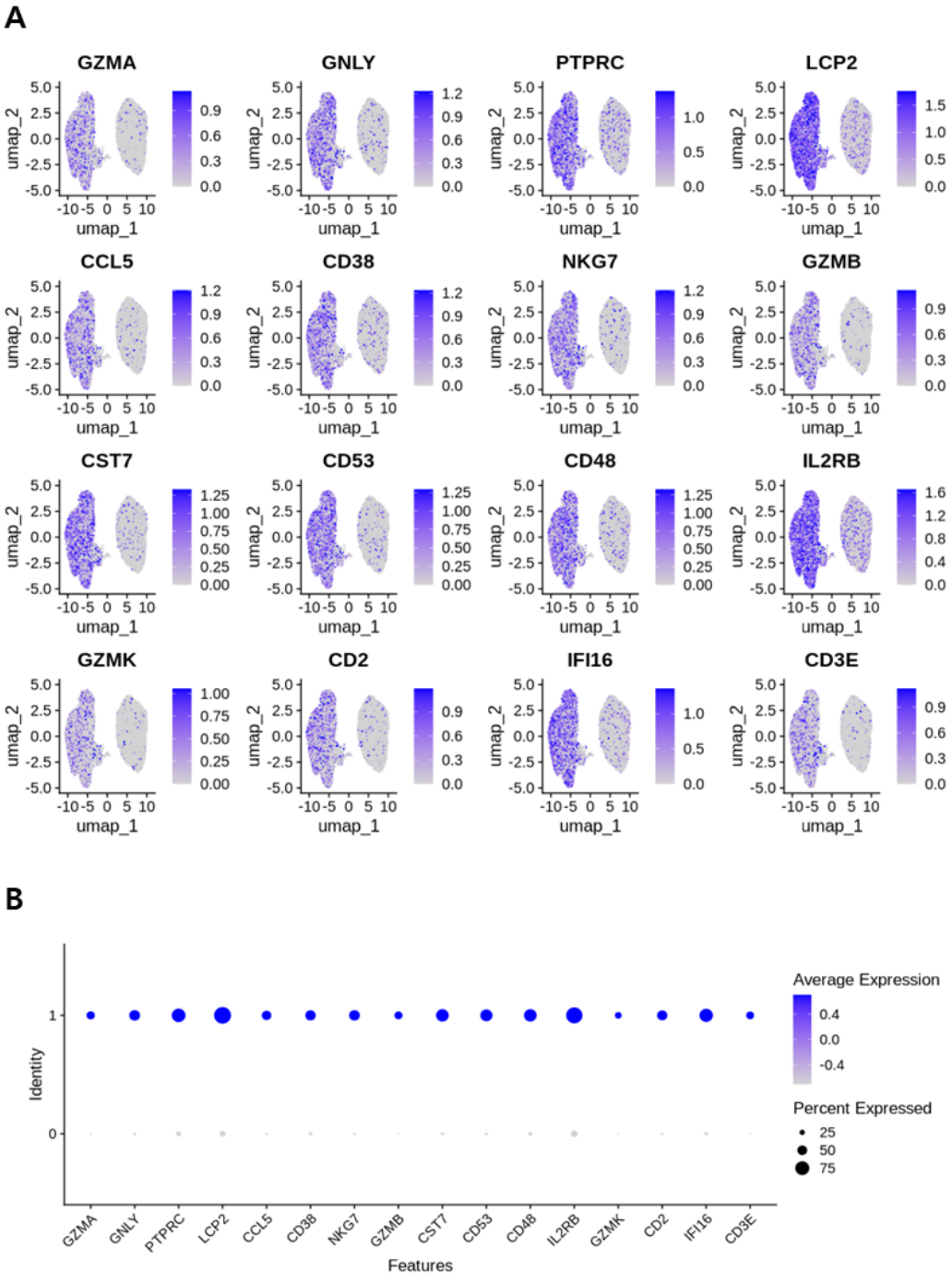




연세대학교
YONSEI UNIVERSITY

Figure 10. Expression of HCT116 cell line signature genes in snRNA-seq clusters. (A) UMAP projection showing the expression patterns of 16 HCT116 signature genes across snRNA-seq clusters. The HCT116 signature genes were identified through bulk mRNA-seq analysis of HCT116_CMV and NK92, selecting genes that are exclusively expressed in the HCT116_CMV cell line and are highly expressed with functional relevance to HCT116. (B) Dot plot showing the fraction of cells (dot size) and mean expression levels (color intensity) of 16 HCT116 signature genes across snRNA-seq clusters.

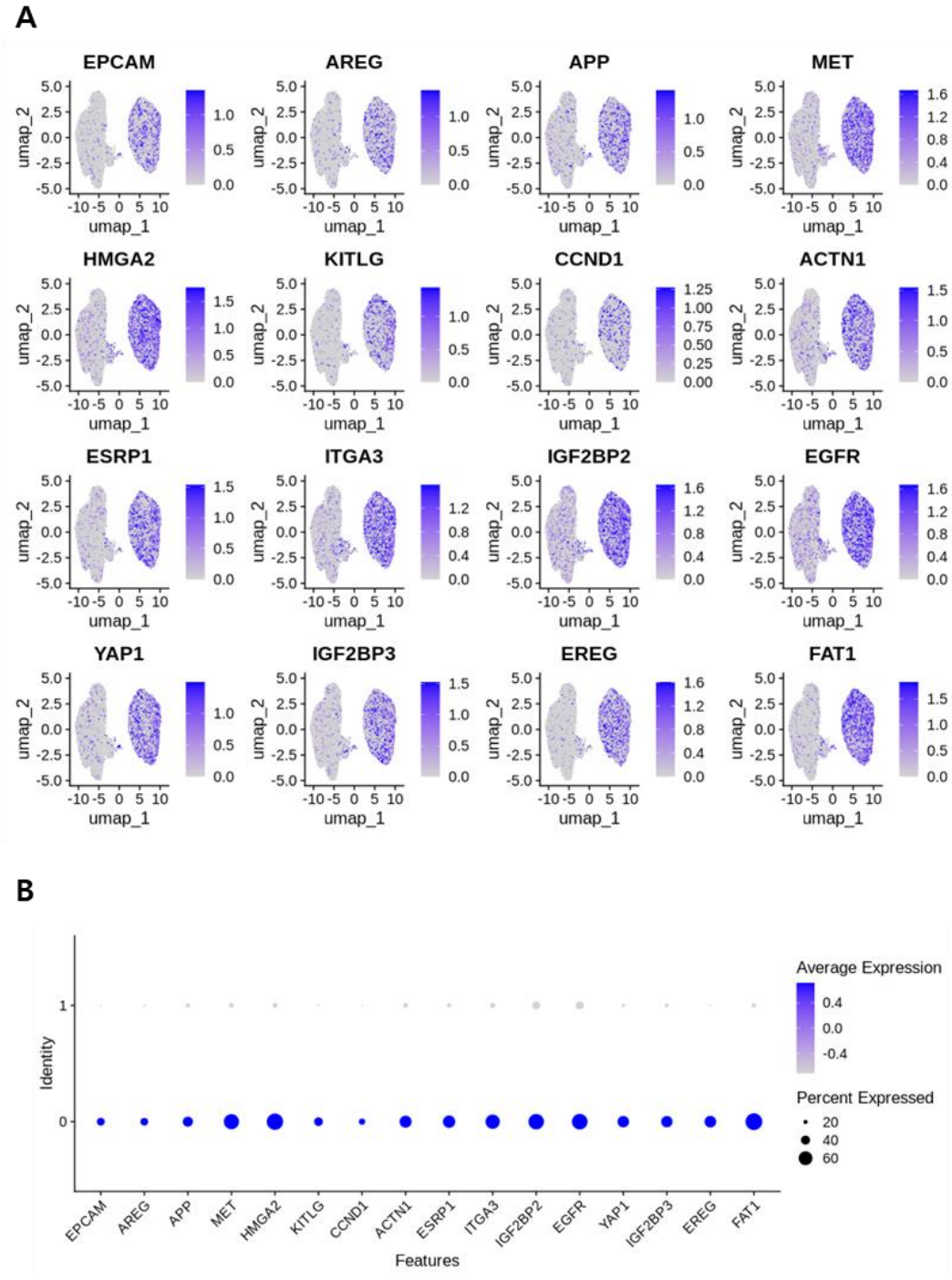
연세대학교
YONSEI UNIVERSITY



연세대학교
YONSEI UNIVERSITY

Figure 11. Gene activity of NK92 cell line signature genes in snATAC-seq clusters. (A) UMAP projection showing the gene activity patterns of 16 NK92 signature genes across snATAC-seq clusters. The NK92 signature genes were identified through bulk ATAC-seq analysis of NK92 and HCT116_CMV, selecting genes that show exclusive gene activity in the NK92 cell line and are highly expressed with functional relevance to NK92. (B) Dot plot showing the fraction of cells (dot size) and average expression levels (color intensity) of 16 NK92 signature genes across snATAC-seq clusters.

연세대학교
YONSEI UNIVERSITY



연세대학교
YONSEI UNIVERSITY

Figure 12. Gene activity of HCT116 cell line signature genes in snATAC-seq clusters. (A) UMAP projection showing the gene activity patterns of 16 HCT116 signature genes across snATAC-seq clusters. The HCT116 signature genes were identified through bulk ATAC-seq analysis of HCT116_CMV and NK92, selecting genes that show exclusive gene activity in the HCT116_CMV cell line and are highly expressed with functional relevance to HCT116. (B) Dot plot showing the fraction of cells (dot size) and average expression levels (color intensity) of 16 HCT116 signature genes across snATAC-seq clusters.

연세대학교
YONSEI UNIVERSITY

3.8. Comparison of cluster-specific IGV profiles with bulk mRNA-seq and ATAC-seq data.

To further validate the identity and characteristics of cell clusters identified through UMAP analysis of snRNA-seq and snATAC-seq data, cluster-specific IGV profiles were compared with IGV profiles derived from bulk mRNA-seq and ATAC-seq data of NK92 and HCT116_CMV cell lines. Bulk mRNA-seq and ATAC-seq data provide an averaged transcriptional landscape and chromatin accessibility across entire cell populations. In this study, bulk mRNA-seq and ATAC-seq profiles were used as references to compare with the cluster-specific IGV profiles generated from snRNA-seq and snATAC-seq data. This comparative analysis confirmed whether the transcriptional and chromatin accessibility profiles of the identified clusters matched the known expression patterns of NK92 and HCT116 cells, thereby enhancing the reliability of cell typing and cluster classification.

From snRNA-seq, Cluster 0 was confirmed to represent the HCT116 cell line, while Cluster 1 represented the NK92 cell line. Additionally, Cluster 2 was identified as an HCT116-derived cluster enriched for rRNA expression. Similarly, from snATAC-seq, Cluster 0 was validated as representing the HCT116 cell line, and Cluster 1 was identified as representing the NK92 cell line.

We first compared IGV profiles for the NK92 signature genes *GNLY* and *GZMA* from NK92 and HCT116 bulk mRNA-seq data with cluster-specific IGV profiles from snRNA-seq. The results revealed pronounced expression of *GNLY* and *GZMA* in Cluster 1, consistent with the signature expression pattern of NK92 cells. In contrast, *GNLY* and *GZMA* expression was minimal in Clusters 0 and 2. These findings strongly suggest that Cluster 1 represents NK92 cells (**Figure 13A**). Next, IGV profiles for *GNLY* and *GZMA*

연세대학교
YONSEI UNIVERSITY

from NK92 and HCT116 bulk ATAC-seq data were compared with cluster-specific IGV profiles from snATAC-seq. Similar to the mRNA-seq data, Cluster 1 exhibited high chromatin accessibility in the promoter and surrounding regions of *GNLY* and *GZMA*. This pattern closely resembled the NK92 bulk ATAC-seq data, indicating that Cluster 1 is associated with NK92 cells. In contrast, Cluster 0 displayed low chromatin accessibility for *GNLY* and *GZMA* (**Figure 13B**).

Next, IGV profiles for the HCT116 signature genes *AREG* and *EPCAM* from NK92 and HCT116 bulk mRNA-seq data were compared with cluster-specific IGV profiles from snRNA-seq. The results demonstrated strong expression of *AREG* and *EPCAM* in Clusters 0 and 2, consistent with the signature expression pattern of HCT116 cells. In contrast, Cluster 1 exhibited minimal expression of these genes, strongly indicating that Clusters 0 and 2 represent HCT116 cells (**Figure 13C**). Furthermore, IGV profiles for *AREG* and *EPCAM* from NK92 and HCT116 bulk ATAC-seq data were compared with cluster-specific IGV profiles from snATAC-seq. Similar to the mRNA-seq results, Clusters 0 exhibited high chromatin accessibility in the promoter and surrounding regions of *AREG* and *EPCAM*, resembling the HCT116 bulk ATAC-seq data. In contrast, Cluster 1 showed low chromatin accessibility for these genes (**Figure 13D**).

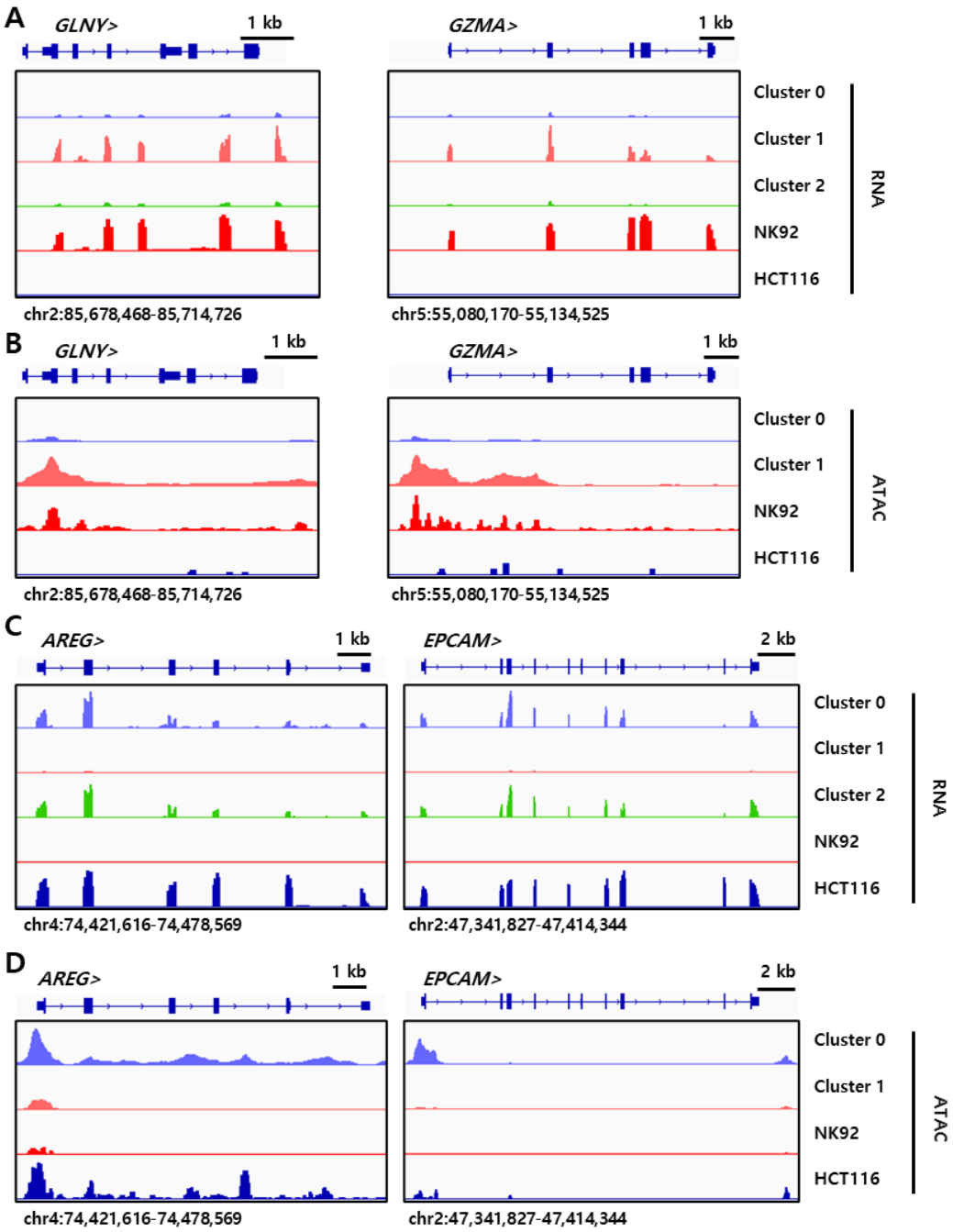
Finally, to distinguish the rRNA-enriched HCT116 cluster in snRNA-seq data, the expression patterns of rRNA genes were analyzed. The results revealed significantly elevated ribosomal RNA (rRNA) expression in Cluster 2, a distinct feature compared to Clusters 0 and 1. These findings clearly demonstrate that Cluster 2 is an rRNA-enriched subcluster derived from HCT116 cells (**Figure 14A**).

In summary, the results confirmed that in snRNA-seq, Cluster 0 represents HCT116, Cluster 1 represents NK92, and Cluster 2 is an rRNA-enriched HCT116 subcluster.

연세대학교
YONSEI UNIVERSITY

Similarly, in snATAC-seq, Cluster 0 represents HCT116, and Cluster 1 represents NK92. These findings demonstrate the utility of RNA expression and chromatin accessibility data in precisely analyzing the differential expression of cell-type-specific genes and chromatin accessibility across different cell types.

연세대학교
YONSEI UNIVERSITY



연세대학교
YONSEI UNIVERSITY

Figure 13. Genome tracks of NK92 and HCT116 signature genes in snRNA-seq and snATAC-seq clusters. (A) IGV tracks visualizing the expression of NK92 signature genes *GNLY* and *GZMA* based on snRNA-seq clusters and bulk mRNA-seq data. (B) IGV tracks showing chromatin accessibility of NK92 signature genes *GNLY* and *GZMA* based on snATAC-seq clusters and bulk ATAC-seq data. (C) IGV tracks visualizing the expression of HCT116 signature genes *AREG* and *EPCAM* based on snRNA-seq clusters and bulk mRNA-seq data. (D) IGV tracks showing chromatin accessibility of HCT116 signature genes *AREG* and *EPCAM* based on snATAC-seq clusters and bulk ATAC-seq data.

연세대학교
YONSEI UNIVERSITY

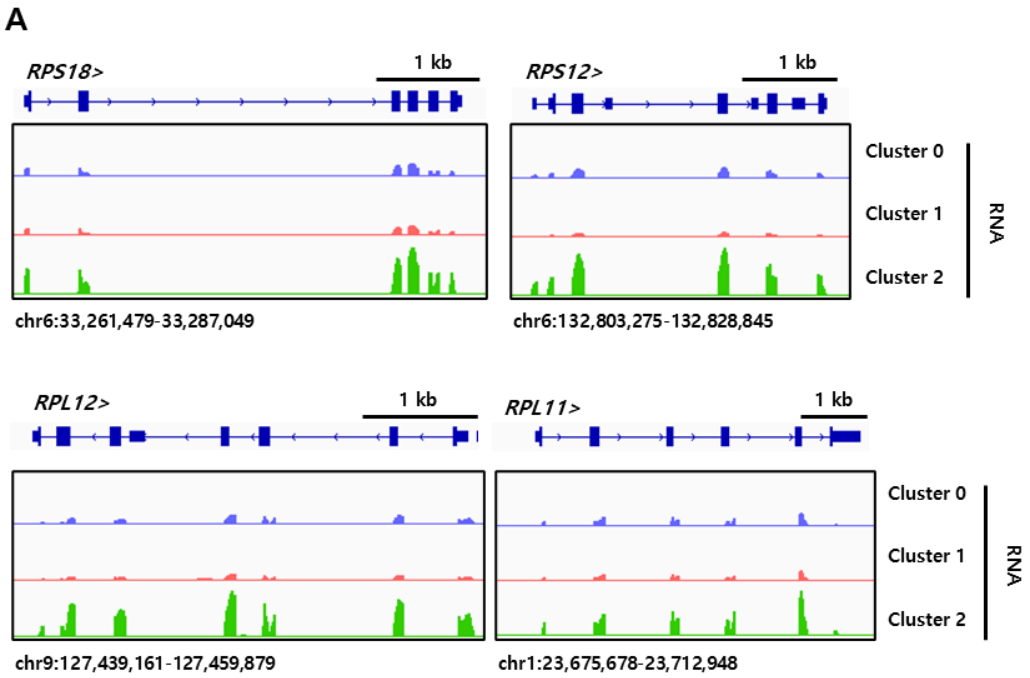


Figure 14. Genome tracks of rRNA genes in snRNA-seq clusters. (A) IGV tracks visualizing the expression of NK92 ribosomal RNA-coding genes *RPS18*, *RPS12*, *RPL12*, and *RPL11* based on snRNA-seq clusters.

연세대학교
YONSEI UNIVERSITY

3.9. Identification of nuclei capable of simultaneously assessing chromatin accessibility and gene expression.

After completing cell typing for each modality (**Figure 15A, 15B**), nuclei with matching barcodes across both modalities were identified. UMAP clustering of each dataset revealed distinct clusters corresponding to the two cell lines. We identified 14,007 nuclei with matching barcodes between the two modalities using high-quality nuclei that passed quality control, including 14,610 snRNA-seq nuclei and 17,833 snATAC-seq nuclei. These nuclei were identified as those containing both transcriptome information and chromatin accessibility data within a single nucleus, enabling the analysis of multimodal data.

To assess whether single nuclei clustered in snRNA-seq and snATAC-seq matched across the two modalities, nuclei sharing the same barcode were visualized by connecting them with lines on the UMAP plots (**Figure 15C**). This analysis revealed that clusters identified as NK92 cells in both modalities exhibited a high degree of barcode matching between the UMAP plots of snRNA-seq and snATAC-seq. However, a portion of the clusters classified as HCT116 cells in snRNA-seq was found to share barcodes with clusters identified as NK92 cells in snATAC-seq (**Figure 15C**). These mismatched nuclei appear to exhibit signals from two different cell types depending on the modality, suggesting that they may result from technical artifacts introduced during the experimental process. Barcode matching analysis can help identify such nuclei, which may introduce bias into multimodal analyses, thereby enabling more accurate interpretation of the data.

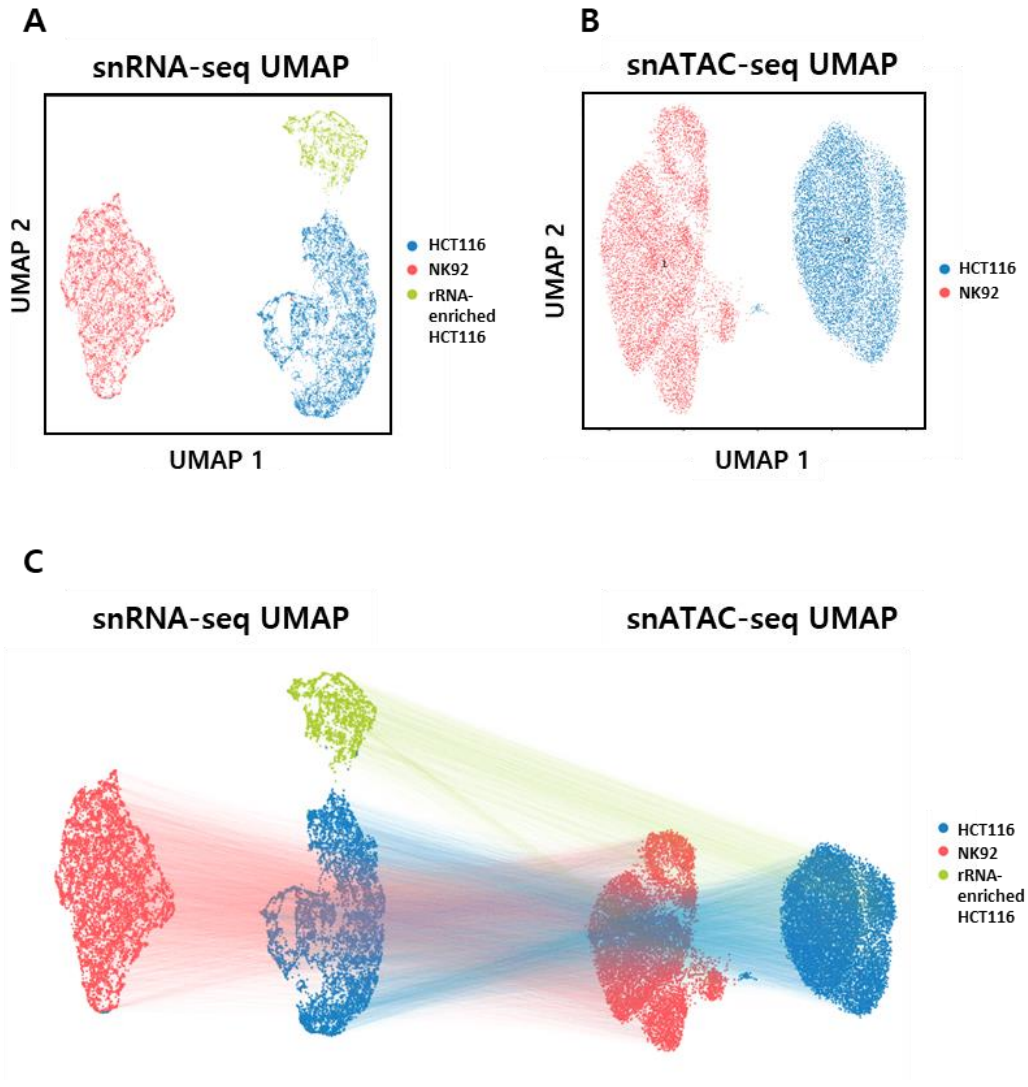
Additionally, we observed that Cluster 2 in the snRNA-seq data, characterized by globally high ribosomal RNA expression, corresponded to the HCT116 cluster in the

연세대학교
YONSEI UNIVERSITY

snATAC-seq data. These findings indicates the presence of a subset of HCT116 cells contaminated with ribosomal RNA within the HCT116 cell cluster identified in the snATAC-seq data.

These findings demonstrate that integrating snRNA-seq and snATAC-seq data enables the precise identification of cell types in mixed samples of different cell lines. Moreover, this integrated analysis compensates for technical errors and enhances the reliability of cell-type classification.

연세대학교
YONSEI UNIVERSITY



연세대학교
YONSEI UNIVERSITY

Figure 15. Barcode matching between snRNA-seq and snATAC-seq UMAP. (A) UMAP visualization of snRNA-seq data showing cluster-specific cell typing of single nuclei derived from a mixed sample of HCT116 and NK92 cell lines. (B) UMAP visualization of snATAC-seq data showing cluster-specific cell typing of single nuclei derived from a mixed sample of HCT116 and NK92 cell lines. (C) UMAP visualization of snRNA-seq and snATAC-seq data, illustrating the connection of cells with identical barcodes.

연세대학교
YONSEI UNIVERSITY

4. DISCUSSION

SHARE-seq is an experimental platform that enables the combined analysis of chromatin accessibility and gene expression at single-cell resolution, providing a cost-effective and highly scalable solution. SHARE-seq enables researchers to elucidate the functional relationships of regulatory elements that control gene expression by leveraging chromatin accessibility and gene expression data. Furthermore, temporal changes in chromatin accessibility and gene expression data can be analyzed to reconstruct cellular lineages and differentiation processes. In this study, SHARE-seq was performed on 20,000 nuclei derived from a mixture of NK92 and HCT116 cells (**Figure 1A**). A portion of the SHARE-seq library was first analyzed through bulk sequencing to confirm that the average gene expression and chromatin accessibility of the mixed cell population could be captured at the bulk level (**Figure 4C, 4D**). Subsequently, single-nucleus sequencing was performed. After quality control, 14,610 nuclei were included in the snRNA-seq dataset (**Figure 5A, 5B**), and 17,833 nuclei were retained in the snATAC-seq dataset (**Figure 6A, 6B**). UMAP visualization was employed to delineate modality-specific clusters, and three distinct clusters were identified in the snRNA-seq data (**Figure 7A**). Cell typing for the snRNA-seq data was conducted by analyzing the top 200 genes uniquely expressed in each cluster (**Figure 7C**) using the Panglao DB_Augmented_2021 and ARCHS4 Cell-lines databases (**Figure 8A, 8B, 8C**). Functional characteristics and pathway enrichment for each cluster were analyzed using the Elsevier Pathway Collection (**Figure 8A, 8B, 8C**). Additionally, differential gene expression (DEG) analysis was performed using bulk mRNA-seq data from NK92 and HCT116 cell lines, identifying genes that were uniquely and highly expressed at the bulk level in each cell line (**Figure 4A**). These genes were classified as

연세대학교
YONSEI UNIVERSITY

cell-type signature genes, and their expression levels across clusters were examined to cell typing (**Figure 9A, 9B, 10A, 10B**). Cell typing of the clusters detected in the snRNA-seq data differentiated the NK92 and HCT116 cell lines and revealed that a subset of HCT116 cells was contaminated with ribosomal RNA. For the snATAC-seq data, UMAP visualization revealed two distinct clusters (**Figure 7B**). Similar to the snRNA-seq analysis, cell typing and pathway enrichment analyses were performed using the top 200 genes with high gene activity scores. Although these analyses did not yield cell-type-specific results for snATAC-seq, the activity of signature genes previously used for snRNA-seq cell typing was examined, revealing differences in gene activity across clusters (**Figure 11A, 11B, 12A, 12B**). Cell typing of the clusters identified in the snATAC-seq data distinguished the NK92 and HCT116 cell lines. To further validate the identity and characteristics of the cell clusters identified through UMAP analysis of snRNA-seq and snATAC-seq data, cluster-specific IGV profiles were compared with IGV profiles derived from bulk mRNA-seq and ATAC-seq data of NK92 and HCT116 cell lines (**Figure 13A, 13B, 13C, 13D, 14A**). This comparative analysis allowed for a precise confirmation of whether the transcriptional and chromatin accessibility profiles of the identified clusters aligned with the previously known expression and chromatin accessibility patterns of NK92 and HCT116 cells.

In summary, snRNA-seq analysis confirmed that Cluster 0 represents HCT116, Cluster 1 represents NK92, and Cluster 2 corresponds to rRNA-enriched HCT116 (**Figure 15A**). Similarly, snATAC-seq analysis validated that Cluster 0 represents HCT116, and Cluster 1 represents NK92 (**Figure 15B**). These findings demonstrate the utility of RNA expression and chromatin accessibility data in accurately analyzing cell-type-specific gene expression and chromatin accessibility differences across distinct cell types.

These analyses enabled the elucidation of chromatin accessibility and gene expression

연세대학교
YONSEI UNIVERSITY

profiles of individual nuclei, which could not be resolved by bulk sequencing. Furthermore, chromatin accessibility and gene expression profiling enabled identifying each nucleus's cell type. After completing cell typing for each modality, barcodes of nuclei identified in snRNA-seq and snATAC-seq analyses were matched (**Figure 15C**). This process yielded a dataset of 14,007 nuclei with jointly profiled gene expression and chromatin accessibility data. This dataset of 14,007 nuclei, containing information from two modalities, can be utilized not only for future clustering analyses across both modalities but also for investigating the functional relationships of regulatory elements controlling gene expression. Additionally, this study validated the reliability of cluster identities by comparing the transcriptomic landscapes and chromatin accessibility patterns of NK92 and HCT116, as previously established in bulk-level experiments, with those of the identified clusters. This comparison confirmed the validity and accuracy of single-nucleus analysis in distinguishing between cell types. By analyzing a mixed sample of NK92 and HCT116 at the single-nucleus level, this study demonstrated that immune and cancer cells can be reliably distinguished using two modalities in future multimodal sequencing studies of the tumor microenvironment (TME). Through this study, high-quality nuclei containing both transcriptome and chromatin accessibility data for each cell type were identified. This study highlights the potential of multimodal data integration for cell-type analysis and suggests its applicability to more complex systems, such as the tumor microenvironment (TME) or tissues with diverse cell types.

연세대학교
YONSEI UNIVERSITY

5. CONCLUSION

In this study, SHARE-seq was used to simultaneously analyze chromatin accessibility and gene expression at the single-nucleus level in a cancer-immune cell mixture composed of NK92 and HCT116 cells. This approach enabled the integrated analysis of gene expression and chromatin accessibility in individual nuclei, demonstrating the capability to accurately distinguish cell types based on these profiles. While bulk sequencing provided only the average gene expression and chromatin accessibility of the mixed cell population, single-cell sequencing revealed detailed information specific to each cell type. Additionally, by comparing IGV profiles of bulk mRNA-seq and ATAC-seq data with those from single-nucleus data, we validated the reliability and accuracy of cluster classification, confirming that the identified clusters aligned with known transcriptomic and chromatin accessibility patterns of NK92 and HCT116 cells (**Figure 16A**). Furthermore, barcodes from both snRNA-seq and snATAC-seq datasets were matched, yielding a high-quality dataset of 14,007 nuclei that integrates information from both modalities.

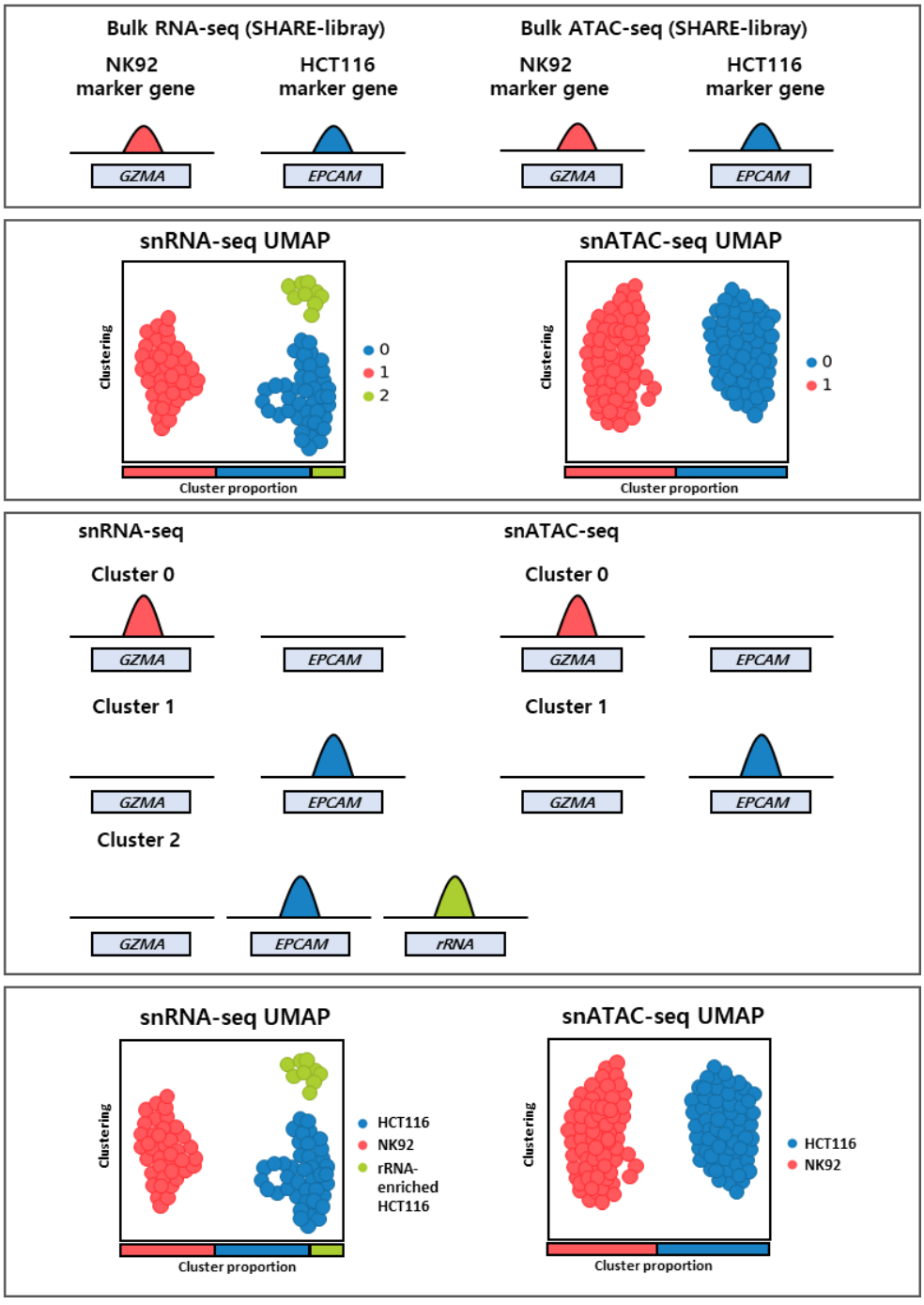
This dataset provides a robust foundation for future clustering analyses and for exploring the functional relationships of regulatory elements controlling gene expression. These findings highlight the utility of single-cell multiomics in resolving cellular heterogeneity and identifying cell-type-specific regulatory mechanisms, which bulk sequencing cannot achieve. By analyzing a mixed sample of NK92 and HCT116 cells at the single-nucleus level, this study demonstrated that immune cells and cancer cells can be reliably distinguished. This approach holds great potential for investigating cell-cell interactions and cellular diversity in complex biological systems, such as the tumor microenvironment (TME). In conclusion, the multimodal single-cell analysis enabled by

연세대학교
YONSEI UNIVERSITY

SHARE-seq offers a powerful tool for unraveling the complexities of diverse biological environments, paving the way for deeper insights into the molecular mechanisms underlying cellular function and disease progression.

연세대학교
YONSEI UNIVERSITY

A



연세대학교
YONSEI UNIVERSITY

Figure 16. cell type identification of single nuclei through two distinct modalities using SHARE-seq. (A) Schematic of cell type identification of single nuclei through two distinct modalities using SHARE-seq.

연세대학교
YONSEI UNIVERSITY

REFERENCES

1. Crick F. Central Dogma of Molecular Biology. Nature 1970;227:561-3.
2. Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. Cell 2007;128:635-8.
3. Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science 1974;184:868-71.
4. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 1997;389:251-60.
5. Morrison O, Thakur J. Molecular Complexes at Euchromatin, Heterochromatin and Centromeric Chromatin. Int J Mol Sci 2021;22.
6. Heitz E. “Das” Heterochromatin der Moose: Bornträger; 1928.
7. Bell O, Burton A, Dean C, Gasser SM, Torres-Padilla M-E. Heterochromatin definition and function. Nature Reviews Molecular Cell Biology 2023;24:691-4.
8. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol 2015;109:21 9 1- 9 9.
9. Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. Nat Protoc 2022;17:1518-52.
10. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 2019;20:207-20.
11. Martin EW, Krietsch J, Reggiardo RE, Sousae R, Kim DH, Forsberg EC. Chromatin accessibility maps provide evidence of multilineage gene priming in hematopoietic stem cells. Epigenetics Chromatin 2021;14:2.

연세대학교
YONSEI UNIVERSITY

12. Baumann C, Zhang X, Zhu L, Fan Y, De La Fuente R. Changes in chromatin accessibility landscape and histone H3 core acetylation during valproic acid-induced differentiation of embryonic stem cells. Epigenetics Chromatin 2021;14:58.
13. Furuichi Y, LaFiandra A, Shatkin AJ. 5'-Terminal structure and mRNA stability. Nature 1977;266:235-9.
14. Darnell JE, Wall R, Tushinski RJ. An adenylic acid-rich sequence in messenger RNA of HeLa cells and its possible relationship to reiterated sites in DNA. Proc Natl Acad Sci U S A 1971;68:1321-5.
15. Edmonds M, Vaughan MH, Jr., Nakazato H. Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. Proc Natl Acad Sci U S A 1971;68:1336-40.
16. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A 1977;74:3171-5.
17. Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell 1977;12:1-8.
18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621-8.
19. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 2008;320:1344-9.
20. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57-63.

연세대학교
YONSEI UNIVERSITY

21. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science 2012;338:1622-6.

22. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015;523:486-90.

23. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun 2019;10:1930.

24. Bendall SC, Simonds EF, Qiu P, Amir el AD, Krutzik PO, Finck R, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 2011;332:687-96.

25. Liu L, Liu C, Quintero A, Wu L, Yuan Y, Wang M, et al. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nat Commun 2019;10:470.

26. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 2018;361:1380-5.

27. Zhu C, Yu M, Huang H, Juric I, Abnousi A, Hu R, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Nat Struct Mol Biol 2019;26:1063-70.

28. Rang FJ, de Luca KL, de Vries SS, Valdes-Quezada C, Boele E, Nguyen PD, et al. Single-cell profiling of transcriptome and histone modifications with EpiDamID. Mol Cell 2022;82:1956-70 e14.

29. Xie Y, Zhu C, Wang Z, Tastemel M, Chang L, Li YE, et al. Droplet-based single-

연세대학교
YONSEI UNIVERSITY

cell joint profiling of histone modifications and transcriptomes. Nat Struct Mol Biol 2023;30:1428-33.

30. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods 2017;14:865-8.

31. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol 2017;35:936-9.

32. Zhang M, Eichhorn SW, Zingg B, Yao Z, Cotter K, Zeng H, et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. Nature 2021;598:137-43.

33. Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nat Rev Mol Cell Biol 2023;24:695-713.

34. 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.

35. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol 2019;37:1452-7.

36. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 2015;348:910-4.

37. Maniyadath B, Zhang Q, Gupta RK, Mandrup S. Adipose tissue at single-cell resolution. Cell Metab 2023;35:386-413.

38. Miranda AMA, Janbandhu V, Maatz H, Kanemaru K, Cranley J, Teichmann SA,

연세대학교
YONSEI UNIVERSITY

et al. Single-cell transcriptomics for the assessment of cardiac disease. Nat Rev Cardiol 2023;20:289-308.

39. 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.

40. Li H, Humphreys BD. Protocol for multimodal profiling of human kidneys with simultaneous high-throughput ATAC and RNA expression with sequencing. STAR Protoc 2024;5:103049.

41. F K. Trim Galore! . 2019.

42. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.

43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.

45. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res 2014;42:W187-91.

46. 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.

47. 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.

48. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.

연세대학교
YONSEI UNIVERSITY

Bioinformatics 2018;34:i884-i90.

49. 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.

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

51. 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.

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

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

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

55. 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.

56. Hardwick SA, Hu W, Joglekar A, Fan L, Collier PG, Foord C, et al. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. Nat Biotechnol 2022;40:1082-92.

57. Lieberman J. The ABCs of granule-mediated cytotoxicity: new weapons in the arsenal. Nat Rev Immunol 2003;3:361-70.

58. Veljkovic Vujaklija D, Dominovic M, Gulic T, Mahmutefendic H, Haller H, Saito S, et al. Granulysin expression and the interplay of granulysin and perforin at the maternal-fetal interface. J Reprod Immunol 2013;97:186-96.

연세대학교
YONSEI UNIVERSITY

59. Bormann F, Stinzing S, Tierling S, Morkel M, Markelova MR, Walter J, et al. Epigenetic regulation of Amphiregulin and Epiregulin in colorectal cancer. Int J Cancer 2019;144:569-81.

60. Lee CC, Yu CJ, Panda SS, Chen KC, Liang KH, Huang WC, et al. Epithelial cell adhesion molecule (EpCAM) regulates HGFR signaling to promote colon cancer progression and metastasis. J Transl Med 2023;21:530.

61. Germain PL, Sonrel A, Robinson MD. pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. Genome Biol 2020;21:227.

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

63. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2018; doi:10.1038/nbt.4314.

64. Franzen O, Gan LM, Bjorkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford) 2019;2019.

65. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun 2018;9:1366.

연세대학교
YONSEI UNIVERSITY

Abstract in Korean

암-면역 세포 혼합체에서 단일 핵 다중유전체 데이터 생산 및 품질 관리

단일 세포에 대한 연구는 세포 간 이질성 탐구, 희귀 세포 유형의 식별, 발달 과정 및 세포 운명에 관한 연구를 가능하게 했다. 그러나 단일 모달리티에 의존한 단일 세포 연구는 세포 내 복잡한 유전자 조절 네트워크에 대한 제한된 정보를 제공한다. 이러한 한계를 극복하기 위해 동일 세포에서 게놈, 후성유전체, 전사체 및 프로테옴을 함께 분석할 수 있는 실험적 방법들이 개발되었다. 현재, 다중유전체학에서 드롭릿 기반의 방법들이 널리 사용되지만, 비용이 높고 처리량이 낮은 단점이 있다. 본 연구에서는 SHARE-seq (Simultaneous High-throughput ATAC and RNA Expression with Sequencing)을 적용하여 면역 세포주인 NK92와 대장암 세포주인 HCT116의 혼합체에서 라이브러리를 생산하고 분석했다. SHARE-seq은 조합 인덱싱 기반의 방법으로, 기존의 드롭릿 기반 다중유전체 기술에 비해 더 높은 처리량과 비용 효율성을 제공한다. 이 방법을 통해 면역세포와 암세포가 혼합된 샘플 내에서 각 세포주에 특이적인 염색질 접근성과 유전자 발현 프로파일을 단일 핵 수준에서 확인했다. 또한 UMAP 분석을 통해 NK92와 HCT116 세포주에 해당하는 뚜렷한 클러스터를 각각의 모달리티에서 구분했다. 마지막으로 snATAC-seq과 snRNA-seq 각각의

연세대학교
YONSEI UNIVERSITY

클러스터에서 일치하는 바코드를 가진 핵들을 확인하였다. 이 핵들은 추후 가중 최근접 이웃(WNN) 분석이나 유전자 발현을 조절하는 조절 요소들의 관계 연구에 활용 가능한 고품질 핵이다.

본 연구는 종양-면역 세포주 혼합 샘플에서 단일 핵 수준으로 염색질 접근성과 유전자 발현을 동시에 분석함으로써, 하나의 핵에서 두 가지의 모달리티를 활용해 면역 세포주와 대장암 세포주를 정밀하게 구분할 수 있음을 입증했다. 또한 향후 염색질 접근성과 유전자 발현 데이터를 통합하여 유전자 발현을 조절하는 조절 요소들의 기능적 관계를 연구할 수 있는 고품질의 핵들을 식별할 수 있음을 보여준다. 이러한 결과는 종양 미세환경(TME)과 같은 복잡한 생물학적 시스템에서 세포 유형을 정밀하게 식별하고, 세포 간 상호작용과 유전자 조절 네트워크를 이해하는 데 기여할 것으로 기대된다. 단일 세포 다중유전체 데이터 통합이 다양한 세포 유형으로 구성된 조직이나 특정 생체 내 환경에서 세포 유형 특성화와 정밀한 분석에 널리 활용될 것으로 기대된다.

핵심되는 말 : 단일 핵, 다중유전체 시퀀싱, 염색질 접근성, 전사체

연세대학교
YONSEI UNIVERSITY

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.