

# **Computational Method for Single Cell ATAC-seq Imputation and Dimensionality Reduction**

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von  
Zhijian Li, M.Sc.  
aus Anyang, China

Berichter: Universitätsprofessor Dr. rer. nat. Thomas Berlage  
Universitätsprofessor Dr. rer. nat. Ivan Gesteira Costa Filho  
Juniorprofessor Michael Thomas Schaub, Ph.D.

Tag der mündlichen Prüfung: 10. August 2022

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.





---

## Abstract (English)

---

Chromatin accessibility, or the physical access to chromatinized DNA, plays an essential role in controlling the temporal and spatial expression of genes in eukaryotic cells. Assay for transposase-accessible chromatin followed by high throughput sequencing (ATAC-seq) is a sensitive and straightforward protocol for profiling chromatin accessibility in a genome-wide manner. Moreover, combined with single-cell sequencing technology, the single-cell ATAC-seq (scATAC-seq) is able to map regulatory variation from hundreds to thousands of cells at single-cell resolution, further expanding its applications.

However, a major drawback of scATAC-seq data is its inherent sparsity. In other words, many open chromatin regions are not detected due to low input or loss of DNA material in the scATAC-seq experiment, leaving a large number of missing values in the derived count matrix. Such a phenomenon is known as “drop-outs” and is also observed in other single-cell sequencing data, such as scRNA-seq. Although many computational methods have been proposed to address this issue for scRNA-seq based on data imputation or denoising, there is a substantial lack of efforts to assess the usability of these methods on scATAC-seq data. Moreover, the development of specific algorithms for imputing or denoising scATAC-seq is still poorly explored yet.

Another critical issue when dealing with the scATAC-seq matrix is the high dimensionality. Because a gene is often regulated by multiple *cis*-regulatory elements (CREs), the number of features in scATAC-seq (i.e., peaks) is usually one order magnitude higher compared with the number of features in scRNA-seq (i.e., genes). This high dimensionality poses a challenge for the analysis of scATAC-seq, such as clustering and visualization. Therefore, it is a common option to first perform dimensionality reduction prior to interpreting the data. However, the standard computational methods for scRNA-seq data are potentially unsuitable for this task due to the low-count information of scATAC-seq data, i.e., a maximum of 2 digestion events is expected for an individual cell in a specific open chromatin region.

In this thesis, we propose scOpen, a computation approach for simultaneous quantification of single-cell open chromatin status and reduction of the dimensionality, to address the aforementioned issues for scATAC-seq data analysis. More formally, scOpen performs imputation and denoising of a scATAC-seq matrix via regularized non-negative matrix factorization (NMF) based on term frequency-inverse document frequency (TF-IDF) transformation. We show that scOpen is able to improve several crucial downstream analysis steps of scATAC-seq data, such as clustering, visualization, *cis*-regulatory DNA interactions and delineation of regulatory features. Moreover, we also demonstrate its power to dissect chromatin accessibility dynamics on large-scale scATAC-seq data from intact mouse kidney tissue. Finally, we perform additional analyses to investigate the regulatory programs that drive the development of kidney fibrosis. Our analyses shed novel light on mecha-

nisms of myofibroblasts differentiation driving kidney fibrosis and chronic kidney disease (CKD). Altogether, these results demonstrate that scOpen is a useful computational approach in biological studies involving single-cell open chromatin data processing.

---

## Abstrakt (Deutsch)

---

Die Zugänglichkeit von Chromatin oder der physikalische Zugang zu chromatinisierter offener DNA spielt eine wesentliche Rolle bei der Kontrolle der zeitlichen und räumlichen Expression von Genen in eukaryontischen Zellen. Der Assay für Transposase-zugängliches Chromatin, gefolgt von Hochdurchsatz-Sequenzierung (ATAC-seq) ist ein sensitives und unkompliziertes Protokoll zur genomweiten Analyse der Chromatinzugänglichkeit. Darüber hinaus ist das Einzelzell-ATAC-seq (scATAC-seq) in Kombination mit der Einzelzell-Sequenzierungstechnologie in der Lage, regulatorische Variationen von Hunderten bis Tausenden von Zellen mit Einzellaufbau abzubilden, was den Anwendungsbereich weiter ausbaut.

Ein großer Nachteil von scATAC-seq-Daten ist jedoch ihre inhärente Datensparsität. Mit anderen Worten, viele offene Chromatinregionen werden aufgrund des geringen Inputs oder des Verlustes von DNA-Material im scATAC-seq-Experiment nicht erkannt, was eine große Anzahl fehlender Werte in der abgeleiteten Zählmatrix hinterlässt. Ein solches Phänomen ist als "Drop-outs" bekannt und wird auch in anderen Einzelzell-Sequenzierungsdaten beobachtet, wie z. B. scRNA-seq. Obwohl viele Computermethoden vorgeschlagen wurden, um dieses Problem für scRNA-seq basierend auf Datenimputation oder Entrauschung anzugehen, gibt es einen erheblichen Mangel an Bemühungen, die Verwendbarkeit dieser Methoden für scATAC-seq-Daten zu bewerten. Darüber hinaus ist die Entwicklung spezifischer Algorithmen zur Imputation oder Entrauschung von scATAC-seq noch weniger erforscht.

Ein weiterer kritischer Punkt beim Umgang mit der scATAC-seq-Matrix ist die hohe Datendimensionalität. Da ein Gen oft durch mehrere cis-regulatorische Elemente (CREs) reguliert wird, ist die Anzahl der Merkmale in scATAC-seq (d.h. Peaks) normalerweise eine Größenordnung höher als die Anzahl der Merkmale in scRNA-seq (d.h. Gene). Diese hohe Dimensionalität stellt eine Herausforderung für die Analyse von scATAC-seq dar, wie beispielsweise Clustering und Visualisierung. Daher ist es eine übliche Option, zuerst eine Dimensionsreduktion durchzuführen, bevor die Daten interpretiert werden. Die Standard-Rechenmethoden für scRNA-seq-Daten sind jedoch aufgrund der geringen Zählung der scATAC-seq-Daten für diese Aufgabe potenziell ungeeignet, d.h. es werden maximal 2 Verdauungsereignisse für eine einzelne Zelle in einer bestimmten offenen Chromatinregion erwartet.

In dieser Dissertation schlage ich scOpen vor, einen Berechnungsansatz zur gleichzeitigen Quantifizierung des offenen Chromatinstatus einzelner Zellen und zur Reduzierung der Dimensionalität, um die oben genannten Probleme für die scATAC-seq-Datenanalyse zu adressieren. Formaler ausgedrückt führt scOpen die Imputation und Rauschunterdrückung einer scATAC-seq-Matrix über eine regularisierte nicht-negative Matrixfaktorisierung (NMF) basierend auf einer Term-Frequenzinversen Dokumentenfrequenz (TF-IDF)-Transformation durch. Ich zeige, dass scOpen mehrere entschei-

dende nachgelagerte Analyseschritte von scATAC-seq-Daten verbessern kann, wie Clustering, Visualisierung, cis-regulatorische DNA-Interaktionen und Abgrenzung regulatorischer Merkmale. Darüber hinaus demonstriere ich seine Leistungsfähigkeit, die Zugänglichkeitsdynamik von Chromatin auf groß angelegten scATAC-seq-Daten aus intaktem Nierengewebe der Maus zu analysieren. Schließlich führen wir zusätzliche Analysen durch, um die regulatorischen Programme zu untersuchen, die die Entwicklung von Nierenfibrose vorantreiben. Unsere Analysen werfen ein neues Licht auf die Mechanismen der Differenzierung von Myofibroblasten, die Nierenfibrose und chronische Nierenerkrankung (CKD) antreiben. Insgesamt zeigen diese Ergebnisse, dass scOpen ein nützlicher rechnerischer Ansatz für biologischen Studien ist, die Einzelzell-Open-Chromatin-Datenverarbeitung beinhalten.

---

## Acknowledgements

---

First and foremost, I would like to thank my supervisor, Prof. Dr. Ivan G. Costa, for providing me with this opportunity to pursue my Ph.D. studies at the Institute for Computational Genomics at RWTH Aachen University. The work presented in this thesis would not be possible without his support.

Furthermore, I would like to acknowledge Prof. Dr. Thomas Berlage and Prof. Dr. Michael Schaub for agreeing to be the referees of my thesis examination procedure as well as Prof. Dr. Martin Grohe for completing the examination committee and Prof. Dr. Holger Hoos for being the chairperson.

Moreover, I wish to thank various people for their contribution to this work: Prof. Dr. Rafael Kramann and Dr. Christoph Kuppe for their advice and assistance on this project; Prof. Dr. Martin Zenke for the help interpreting results; Dr. Susanne Ziegler, Dr. Nazanin Kabgani, and Dr. Sylvia Menzel for the laboratory work.

I would also like to extend my thanks to my colleagues, who helped me some of the analysis and proofreading in this study or with productive discussions: Dr. Chao-Chung Kuo, M.Sc. Tiago Maie, M.Sc. Mingbo Cheng, and M.Sc. James Nagai.

I want to thank my family and friends for all the support given during my stay in Aachen. Especially, I am very grateful to my parents who give me the best possible start in life and support me in every situation. Finally, I want to thank my girlfriend Ping for all her support.



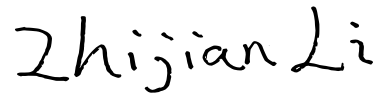
---

## Selbstständigkeitserklärung

---

Ich versichere hiermit an Eides statt, daß ich die vorliegende Doktorarbeit selbstständig und ohne unzuläßige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich habe die Grundsätze zur Sicherung guter wissenschaftlicher Praxis der RWTH Aachen zur Kenntnis genommen und eingehalten.

Aachen, August 12, 2022



---

(Zhijian Li)





---

## Publications

---

As required by §5(3) of the

*Promotionsordnung für die Fakultät für Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfälischen Technischen Hochschule Aachen vom 27.09.2010 in der Fassung der zweiten Ordnung zur Änderung der Promotionsordnung vom 30.06.2014 (veröffentlicht als Gesamtfassung),*

a declaration of results that are published by the author as well as particular contributions to co-authored publications follows.

I am the main author of the following publication, which is co-authored with Prof. Dr. Ivan G. Costa and our experimental collaborators, Prof. Dr. Rafael Kramann, and Dr. med. Christoph Kuppe from the Institute of Experimental Medicine and System Biology in RWTH Aachen University Medical School. All the results presented in this thesis were published in this publication:

- **Z. Li\***, C. Kuppe\*, S. Ziegler, M. Cheng, N. Kabgani, S. Menzel, M. Zenke, R. Kramann, I. G Costa. Chromatin-accessibility estimation of single-cell ATAC data with scOpen. **Nature Communications** 12, 1 (2021)

I am also the main author of the following publication, which is co-authored with Prof. Dr. Ivan G. Costa and our experimental collaborators, Prof. Dr. Martin Zenke, and M.Sc. Thomas Look from the Department of Cell Biology in RWTH Aachen University Medical School.

- **Z. Li**, M. H. Schulz, T. Look, M. Begemann, M. Zenke, I. G. Costa. Identification of transcription factor binding sites using ATAC-seq. **Genome Biology** 20, 45 (2019).

I developed HINT-ATAC, a computational method based on hidden Markov model to identify transcription factor binding sites using bulk ATAC-seq data. I also proposed differential footprinting analysis to compare the transcription factor activity between different biological conditions. The method developed in this publication is adapted in this thesis to analyze single cell ATAC-seq.

Furthermore, I am also the co-first author of the following publication:

- C. Kuppe\*, R. Flores\*, **Z. Li\***, S. Hayat, M. Hannani, J. Tanevski, M. Halder, M. Cheng,

S. Ziegler, X. Zhang, F. Preisker, N. Kaesler, Y. Xu, R. M. Hoogenboezem, E. M. Bindels, R. K. Schneider, H. Milting, I. G. Costa, J. S. Rodriguez, R. Kramann. Spatial multi-omic map of human myocardial infarction. **Nature** (2022).

Kuppe, et al. used single-cell gene expression, chromatin accessibility and spatial transcriptome for profiling of different physiological zones and time points of human myocardial infarction and human control myocardium to generate an integrative high-resolution map of cardiac remodeling. The method proposed in this thesis was used to perform dimensionality reduction for single cell ATAC-seq data. I also carried out primarily all single cell ATAC-seq data analysis.

Finally, the following publications did not directly contribute to this thesis, but shaped a general understanding of next-generation sequencing analysis, single-cell analysis, transcription factor motif analysis and machine learning. They were published during my Ph.D. studies in RWTH Aachen University.

- M. Cheng, **Z. Li**, I.G. Costa. MOJITOO: a fast and universal method for integration of multimodal single-cell data. **Bioinformatics** 38, (2022)
- A. Philippi, S. Heller, I.G. Costa, V. Senée, M. Breunig, **Z. Li**, G. Kwon, R. Russell, A. Illing, Q. Lin, M. Hohwieler, A. Degavre, P. Zalloua, S. Liebau, M. Schuster, J. Krumm, X. Zhang, R. Geusz, J.R. Benthuyzen, A. Wang, J. Chiou, K. Gaulton, H. Neubauer, E. Simon, T. Klein, M. Wagner, G. Nair, C. Besse, C. Dandine-Roulland, R. Olaso, J. Deleuze, B. Kuster, M. Hebrok, T. Seufferlein, M. Sander, B.O. Boehm, F. Oswald, M. Nicolino, C. Julier, A. Kleger. Mutations and variants of ONECUT1 in diabetes. **Nature Medicine** 27, 11 (2021)
- S. Heller, **Z. Li**, Q. Lin, R. Geusz, M. Breunig, M. Hohwieler, X. Zhang, G. Nair, T. Seufferlein, M. Hebrok, M. Sander, C. Julier, A. Kleger\*, I. G. Costa\*. Transcriptional and open chromatin landscape during pancreatic differentiation reveals a role of ONECUT1 in inducing the endocrine transcriptional program. **Communications Biology** 4, 1 (2021)

I implemented scOpen and built the benchmarking dataset. All figures and tables in this thesis are authored by me and exceptions were explicitly stated in their respective captions. I performed all computational analyses in this study with the aid of Prof. Dr. Ivan G. Costa. All chapters of this thesis have been written by me. Prof. Dr. Ivan G. Costa provided support in all stages of this research including thesis manuscript writing. Mostly out of habit and to honor the fact that research is rarely an entirely solitary process I will, in this thesis, rely on the use of the first-person plural pronoun “we” in the text, as a nosism.

---

## List of Figures

---

1.1	Comparison of bulk and single-cell ATAC-seq . . . . .	2
1.2	Thesis overview . . . . .	3
1.3	A novel computation method for scATAC-seq data imputation and dimensionality reduction . . . . .	4
2.1	Chromatin organization . . . . .	8
2.2	Chromatin accessibility dynamics . . . . .	9
2.3	Specificity of TF-DNA interaction . . . . .	10
2.4	Measuring chromatin accessibility using ATAC-seq . . . . .	11
2.5	Measuring chromatin accessibility using single cell ATAC-seq . . . . .	12
2.6	Droplet-based scATAC-seq . . . . .	13
2.7	Computational analysis of bulk ATAC-seq . . . . .	14
2.8	Analytic workflow of scATAC-seq . . . . .	16
2.9	Statistical comparison of scATAC-seq and scRNA-seq data . . . . .	20
2.10	Computational methods for single-cell data denoising and imputation . . . . .	24
3.1	A toy example of scATAC-seq data normalization . . . . .	33
3.2	Conceptual illustration of the defined NMF model . . . . .	36
3.3	Convergence curve of the model with different initialization approaches . . . . .	38
3.4	Estimation of the number of components using elbow detection algorithm . . . . .	40
3.5	Workflow of scOpen . . . . .	41
4.1	Experimental design for evaluation of the imputation methods . . . . .	53
4.2	An example of the Precision-Recall curves . . . . .	54
4.3	Experimental design for evaluation of the dimensionality reduction methods . . . . .	56
4.4	Experimental design for evaluation of the downstream analysis methods . . . . .	56
4.5	Co-accessibility prediction and evaluation . . . . .	57
5.1	Estimation of the number of components for the simulation data . . . . .	66
5.2	Evaluation of the number of components using the simulation data . . . . .	67
5.3	Evaluation of the regularization parameter using the simulation data . . . . .	68
5.4	Evaluation of the memory and running time requirements for imputation methods . . . . .	69
5.5	Evaluation of the imputation methods based on imputation accuracy . . . . .	70
5.6	Association between the number of cells and AUPR . . . . .	71
5.7	Evaluation of the imputation methods based on distance accuracy . . . . .	72

5.8	Evaluation of the imputation methods based on clustering accuracy . . . . .	73
5.9	Visualization of imputation methods on benchmarking datasets . . . . .	74
5.10	Visualization of imputation methods on benchmarking datasets . . . . .	75
5.11	Evaluation of the dimensionality reduction methods using distance accuracy . . . . .	76
5.12	Evaluation of the dimensionality reduction methods using clustering accuracy . . . . .	77
5.13	Evaluation of the downstream analysis methods . . . . .	78
5.14	Visualization of the downstream analysis methods . . . . .	79
5.15	Evaluation of Cicero predicted DNA regulatory elements interactions . . . . .	80
5.16	Visualization of Cicero predicted DNA regulatory elements interactions . . . . .	81
5.17	Visualization of co-accessibility score . . . . .	82
5.18	Evaluation of batch correction using different dimensionality reduction methods . . . . .	83
5.19	Annotation of Uuo scATAC-seq data . . . . .	83
5.20	Visualization of TF activity across different cell types and time points . . . . .	84
5.21	Visualization of myofibroblast differentiation . . . . .	85
5.22	Identification and validation of Runx1 for myofibroblast differentiation . . . . .	86
5.23	Overexpression of Runx1 . . . . .	87
5.24	Prediction of Runx1 target genes . . . . .	88
5.25	Tgfb1 is regulated by Runx1 in myofibroblast . . . . .	89
A.1	Overall rank of the imputation methods . . . . .	105
A.2	The combined rank of the dimensionality reduction methods based on distance and clustering accuracy . . . . .	117
A.3	Evaluation of the predicted interactions by Cicero . . . . .	118
A.4	Visualization of co-accessibility score . . . . .	119
A.5	Evaluation of the predicted interactions by Cicero using down-sampled data . . . . .	120
A.6	Quality control of Uuo scATAC-seq data . . . . .	121
A.7	Visualization of integrated data by using different dimensionality reduction methods . . . . .	122
A.8	Visualization of marker genes for each annotated cell type . . . . .	123
A.9	Visualization of annotated cell types for each time point . . . . .	124
A.10	Twist2 is regulated by Runx1 in myofibroblast . . . . .	125

---

## List of Tables

---

2.1	Overview of computational denoising and imputation methods . . . . .	25
2.2	Overview of computational methods for scATAC-seq dimensionality reduction . . . . .	27
3.1	scOpen tool python package dependencies . . . . .	42
3.2	Methodological comparison between scOpen and other methods . . . . .	43
4.1	Statistics of the benchmarking datasets used in this thesis . . . . .	49
A.1	Statistics of the positive and negative peaks in each benchmarking dataset . . . . .	106
A.2	Friedman-Nemenyi test of the number of components based on imputation accuracy . . . . .	107
A.3	Friedman-Nemenyi test of the number of components based on clustering accuracy . . . . .	108
A.4	Friedman-Nemenyi test of the regularization based on imputation accuracy . . . . .	109
A.5	Friedman-Nemenyi test of the regularization parameters based on clustering accuracy . . . . .	109
A.6	Friedman-Nemenyi test of the imputation accuracy for Cell line dataset . . . . .	109
A.7	Friedman-Nemenyi test of the imputation accuracy for Hematopoiesis dataset . . . . .	110
A.8	Friedman-Nemenyi test of the imputation accuracy for T cells dataset . . . . .	110
A.9	Friedman-Nemenyi test of the imputation accuracy for PBMC dataset . . . . .	111
A.10	Friedman-Nemenyi test of the distance accuracy for Cell line dataset . . . . .	111
A.11	Friedman-Nemenyi test of the distance accuracy for Hematopoiesis dataset . . . . .	112
A.12	Friedman-Nemenyi test of the distance accuracy for T cells dataset . . . . .	112
A.13	Friedman-Nemenyi test of the distance accuracy for PBMC dataset . . . . .	113
A.14	Friedman-Nemenyi test of the clustering accuracy for Cell line dataset . . . . .	113
A.15	Friedman-Nemenyi test of the clustering accuracy for Hematopoiesis dataset . . . . .	114
A.16	Friedman-Nemenyi test of the clustering accuracy for T cells dataset . . . . .	114
A.17	Friedman-Nemenyi test of the clustering accuracy for PBMC dataset . . . . .	115
A.18	Friedman-Nemenyi test of the distance accuracy for dimensionality reduction methods and Cell line dataset . . . . .	115
A.19	Friedman-Nemenyi test of the distance accuracy for dimensionality reduction methods and Hematopoiesis dataset . . . . .	115
A.20	Friedman-Nemenyi test of the distance accuracy for dimensionality reduction methods and T cells dataset . . . . .	116
A.21	Friedman-Nemenyi test of the distance accuracy for dimensionality reduction methods and PBMC dataset . . . . .	116



---

## Glossary

---

<b>ARI</b>	Adjusted rand index
<b>ATAC-seq</b>	Assay for transposase-accessible chromatin using sequencing
<b>AUPR</b>	Area under the precision recall curve
<b>BAM</b>	Binary alignment map
<b>BED</b>	Browser extensible data
<b>bp</b>	Base pair
<b>ChIP-seq</b>	Chromatin immunoprecipitation with sequencing.
<b>CRE</b>	<i>Cis</i> -regulatory element
<b>DCA</b>	Deep count autoencoder
<b>DNA</b>	Deoxyribonucleic acid
<b>ELBO</b>	Evidence lower bound
<b>EM</b>	Expectation-maximization
<b>FRiP</b>	Fraction of reads in peaks
<b>GEM</b>	Gel bead in emulsion
<b>GEO</b>	Gene expression omnibus
<b>KNN</b>	K-nearest neighbor
<b>LDA</b>	Latent dirichlet allocation
<b>LSI</b>	Latent semantic indexing
<b>MAGIC</b>	Markov affinity-based graph imputation of cells
<b>nm</b>	Nanometre
<b>NMF</b>	Non-negative matrix factorization
<b>NNDSVD</b>	Non-negative double singular value decomposition
<b>PCA</b>	Principal component analysis
<b>PCR</b>	Polymerase chain reaction
<b>SAVER</b>	Single-cell analysis Via expression recovery
<b>scABC</b>	Single cell accessibility based clustering

<b>SCALE</b>	Single-cell ATAC-seq analysis via Latent feature Extraction
<b>scBFA</b>	Single-cell binary factor analysis
<b>scRNA-seq</b>	Single-cell RNA sequencing
<b>SNR</b>	Signal-to-noise ratio
<b>SVD</b>	Singular value decomposition
<b>t-SNE</b>	t-distributed stochastic neighbor embedding
<b>TF</b>	Transcription factor
<b>TF-IDF</b>	Term frequency–inverse document frequency
<b>TFBSs</b>	Transcription factor binding sites
<b>TSS</b>	Transcription start site
<b>UMAP</b>	Uniform manifold approximation and projection
<b>UO</b>	Unilateral Ureter Obstruction
<b>VAE</b>	Variational autoencoder
<b>ZINB</b>	Zero-inflation negative binomial



---

## Contents

---

Abstract (English) . . . . .	ii
Abstrakt (Deutsch) . . . . .	iv
Acknowledgements . . . . .	v
Selbstständigkeitserklärung (Declaration of Authenticity) . . . . .	vii
Publications . . . . .	x
List of Figures . . . . .	xii
List of Tables . . . . .	xiii
Glossary . . . . .	xv
Glossary . . . . .	xvi
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Contributions . . . . .	2
1.3 Thesis Structure . . . . .	5
<b>2 Background</b>	<b>7</b>
2.1 DNA Organization, Chromatin Accessibility, and Gene Regulation . . . . .	7
2.1.1 DNA Organization . . . . .	7
2.1.2 Chromatin Accessibility and Gene Regulation . . . . .	8
2.2 Profiling Chromatin Accessibility with ATAC-seq . . . . .	9
2.2.1 Bulk ATAC-seq . . . . .	9
2.2.2 Single Cell ATAC-seq . . . . .	10
2.3 Computational Analysis of ATAC-seq . . . . .	12
2.3.1 Bulk ATAC-seq . . . . .	13
2.3.2 Single Cell ATAC-seq . . . . .	15
2.4 Related Works . . . . .	18
2.4.1 Challenges of Analysis scATAC-seq . . . . .	19
2.4.2 Computational Methods for Single Cell Data Denoising and Imputation . . .	19
2.4.3 Computational Methods for scATAC-seq Dimensionality Reduction . . . . .	25
2.5 Discussion . . . . .	27
<b>3 Methods</b>	<b>31</b>
3.1 Notation . . . . .	31
3.2 Data Normalization . . . . .	32
3.2.1 Binarization . . . . .	32

3.2.2	Transformation . . . . .	32
3.3	Data Imputation and Dimensionality Reduction . . . . .	34
3.3.1	NMF . . . . .	34
3.3.2	Solving the Optimization Problem . . . . .	35
3.3.3	Initialization . . . . .	37
3.3.4	Determining the Hyper-parameters . . . . .	39
3.4	Implementation . . . . .	40
3.5	Discussion . . . . .	42
<b>4</b>	<b>Experiments</b>	<b>45</b>
4.1	Technical Validation . . . . .	45
4.1.1	Data . . . . .	45
4.1.2	scOpen Parameter Selection . . . . .	49
4.1.3	Execution of Computational Methods . . . . .	49
4.1.4	Evaluation of Computational Methods . . . . .	52
4.1.5	Statistical Methods . . . . .	57
4.2	Biological Validation . . . . .	58
4.2.1	Applying scOpen to scATAC-seq Data from Complex Disease . . . . .	58
4.2.2	Characterizing Gene Regulation During Myofibroblast Differentiation . . . . .	60
4.3	Discussion . . . . .	62
<b>5</b>	<b>Results</b>	<b>65</b>
5.1	Technical Validation . . . . .	65
5.1.1	scOpen Parameter Selection . . . . .	65
5.1.2	Evaluation of Imputation Methods . . . . .	66
5.1.3	Evaluation of Dimensionality Reduction Methods . . . . .	71
5.1.4	Evaluation of Downstream Analysis Methods . . . . .	75
5.1.5	Evaluation of Co-accessibility Analysis . . . . .	76
5.2	Biological Validation . . . . .	77
5.2.1	Applying scOpen to Complex Disease scATAC-seq Data . . . . .	77
5.2.2	Characterizing Gene Regulation During Myofibroblast Differentiation . . . . .	83
5.2.3	Identification and Validation of Runx1 Target Genes . . . . .	88
5.3	Discussion . . . . .	90
<b>6</b>	<b>Discussion and Conclusion</b>	<b>91</b>
6.1	Discussion . . . . .	91
6.2	Conclusion and Future Work . . . . .	93
<b>A</b>	<b>Appendix</b>	<b>105</b>
A.1	Technical Validation . . . . .	105
A.2	Biological Validation . . . . .	121

---

## Introduction

---

### 1.1 Motivation

---

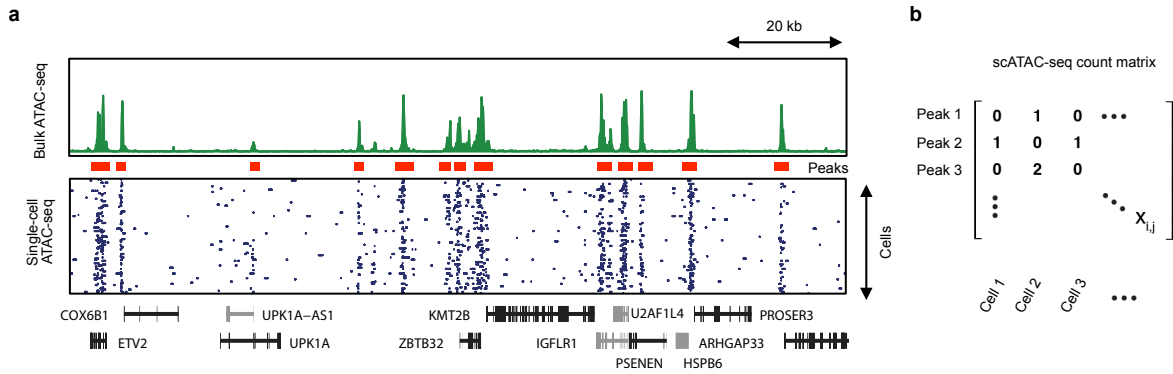
All known living organisms are composed of cells. In a multicellular organism, various cell types are often specialized to perform a unique and specific function. These cell types express very different sets of genes, despite that they carry the exact genetic instructions encoded in the form of deoxyribonucleic acid (DNA) molecules. For example, it is estimated that there are approximately 200 different types of cells in a human body, and each cell contains the same genome. Naturally, a question is how the same genome gives rise to so many kinds of cell types in the human body. The answer is that each cell can only use a proportion of genes, which are controlled by regulatory features and chromatin accessibility. Chromatin is a complex of DNA and proteins, and it exists in two distinct states, i.e., open versus closed. The open chromatin has a loose structure and the genes in these regions are expressed, while the closed chromatin has a condensed structure and the genes in these regions are inactive. Therefore, it is essential to identify the accessible DNA regions in a cell type to understand the molecular mechanism of its gene expression pattern and cell identity.

ATAC-seq (Assay for Transposase-Accessible Chromatin followed by sequencing) is a sensitive and straightforward protocol for profiling chromatin accessibility in a genome-wide and high-throughput manner (Buenrostro et al., 2013, 2015a). It has been successfully applied to investigate the chromatin status during human blood cellular differentiation (Lara-Astiaso et al., 2014), define the chromatin accessibility landscape of primary human cancer (Corces et al., 2018), create an atlas of open chromatin for mouse immune system (Yoshida et al., 2019), among others. Moreover, careful consideration of digestion events by the enzyme Tn5 allows insights on regulatory elements, such as positions of nucleosomes (Buenrostro et al., 2013; Schep et al., 2015), transcription factor binding sites, and the activity level of transcription factors (Li et al., 2019).

Traditionally, ATAC-seq takes 500 to 50,000 cells as input and generates an average chromatin accessibility profile, thus obscuring the biological differences between individual cells. However, cell-to-cell variation is a universal feature of life that affects a wide range of biological phenomena (Buenrostro et al., 2015b). Moreover, technical advances have allowed for characterization of the transcriptome at single-cell resolution (scRNA-seq) (Tang et al., 2009; Islam et al., 2011; Hashimshony et al., 2012; Klein et al., 2015). By combining ATAC-seq and single-cell sequencing techniques, single-cell ATAC-seq (scATAC-seq) was developed to allow to measure chromatin states at single-cell resolution (Buenrostro et al., 2015b; Cusanovich et al., 2015). Figure 1.1a shows an example of bulk and single-cell ATAC-seq data from the same genomic coordinates in GM12878 cells. Since

## 1.2. Contributions

then, scATAC-seq has been dramatically improved by reducing the cost and increasing the throughput of the assay (Chen et al., 2018; Satpathy et al., 2019; Lareau et al., 2019).



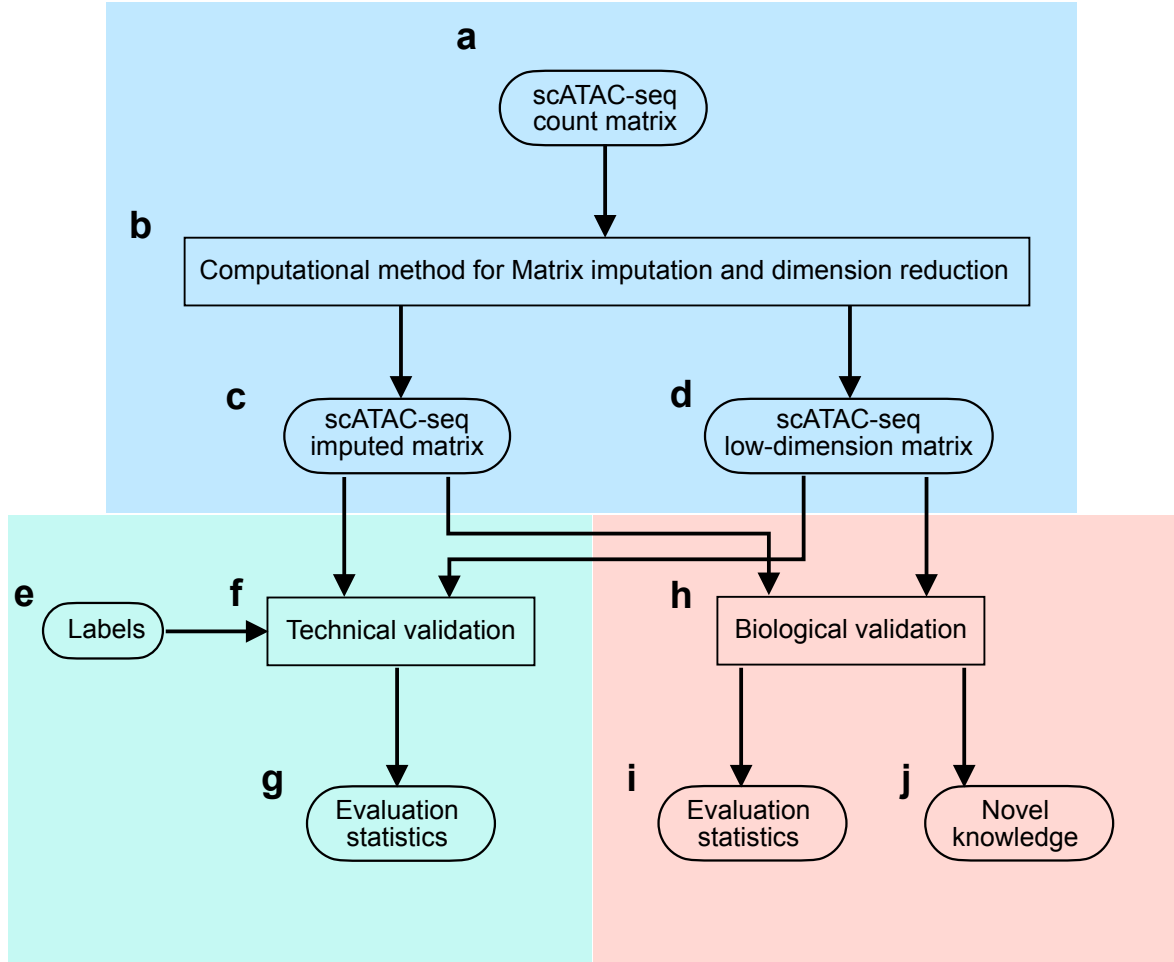
**Figure 1.1: Comparison of bulk and single-cell ATAC-seq.** **a**, Genomic tracks showing bulk (top) and single-cell ATAC-seq (bottom) profiles from GM12878 B lymphoblasts. The red bars in the middle represent open chromatin regions (i.e., peaks). For the bulk ATAC-seq, the y-axis represents the average chromatin accessibility of GM12878 cells. For the scATAC-seq, profiles are obtained from 100 cells. *Source: Satpathy et al. (2019)* (modified to fit thesis format and/or clarify key points). **b**, A matrix representing the scATAC-seq data for downstream analysis. Each column represents a cell and each row represents a peak as shown in **a**. Each entry in the matrix represent the number of reads observed in each peak and cell.

Computational methods are crucial for scATAC-seq data analysis, and the rapid development of scATAC-seq techniques has introduced unprecedented challenges in this regard. After constructing a count matrix from scATAC-seq data (Figure 1.1b), one challenge is how to deal with sparsity, i.e., most of the elements in the matrix are zero. Notably, the sparsity is intrinsic to the scATAC-seq data since even in high-quality experiment, most accessible regions are not transposed due to Tn5 transposition efficiency, leading to many loci to have zero detected alleles (Granja et al., 2021). Another challenge is high dimensionality. The features of scATAC-seq data are defined by performing peak calling based on aggregate data, and the number of peaks is often remarkably high (usually  $10^6$  peaks). Finally, scATAC-seq data has low information content for each peak in individual cells, given that a particular site only has two alleles in a single cell. Altogether, a computational method that can deal with the sparsity, high-dimensionality and low information content of scATAC-seq data is clearly needed.

## 1.2 Contributions

In this thesis, we: (1) present a novel computational method to impute the missing values and reduce the dimensions of single-cell open chromatin data, (2) technically benchmark the performance of our method using both simulated and real-world scATAC-seq datasets for matrix imputation and dimensionality reduction, (3) biologically validate our method by applying it to a large-scale scATAC-seq

data generated from whole mouse kidneys. Figure 1.2 presents an overview of this thesis.



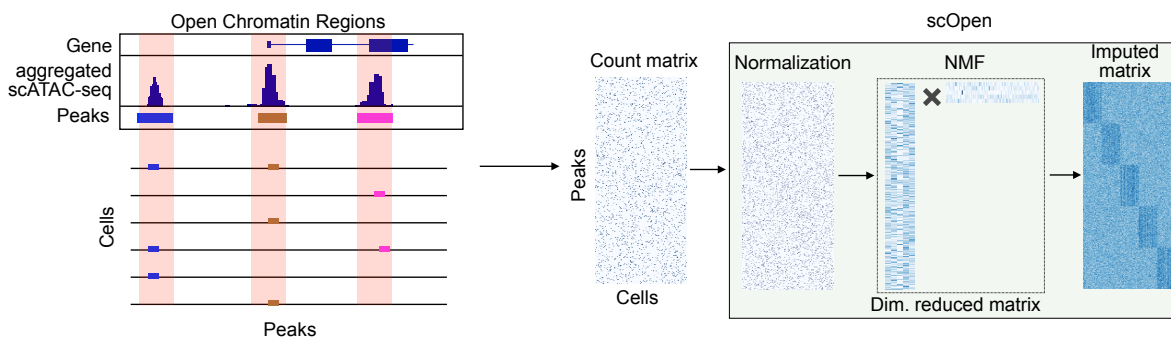
**Figure 1.2: Thesis overview.** This figure shows the workflow of the thesis. Boxes with round-shaped edges represent data, and square-shaped edges describe computational methods. The proposed computation method is highlighted with blue, technical validation with green and biological validation with red.

More specifically, we listed our contributions as follows:

- **A Novel computational method for scATAC-seq data imputation and dimensionality reduction:** We derived a computational approach based on term frequency-inverse document frequency (TF-IDF) transformation and non-negative matrix factorization (NMF) to impute the missing values and reduce the dimensions for a given scATAC-seq count matrix. The TF-IDF transformation normalizes the data to appropriately reflect how important a peak is to a particular cell. Given the non-negativity of TF-IDF transformed matrix, it is a natural choice to use NMF to perform imputation and dimensionality reduction. The experiments have shown that our method can estimate single cell chromatin accessibility status accurately by data imputation and provide a better low-dimensional representation for downstream analysis than its competitors. (Figure 1.3) depicts the overall workflow of the proposed method.

## 1.2. Contributions

- **A Comprehensive approach for simulating scATAC-seq data:** To evaluate the hyper-parameters in our method, we extended the simulation process proposed by Chen et al. (2019) to generate a simulated scATAC-seq dataset. This new method used a negative binomial distribution to model the number of fragments per cell. It also introduced a parameter to control the fraction of reads in peaks (FRiP), thus providing a better simulation for real-world scATAC-seq data.
- **A Novel strategy for evaluating scATAC-seq imputation methods:** We also proposed a novel technique to directly access the imputation results. Specifically, we defined the true label for each peak based on the corresponding bulk data and estimated how many correct peaks become non-zero and vice-versa. Moreover, we ranked these by the chromatin accessibility scores provided by imputation methods and calculated the area under the precision-recall curve (AUPR). This evaluation has the advantage of being independent of downstream analysis, i.e., clustering, and was used to compare our method to the competing methods.
- **A comprehensive evaluation of scATAC-seq imputation methods:** We performed a thorough evaluation for scATAC-seq imputation methods, including: (1) our novel approach; (2) five state-of-the-art imputation methods for scRNA-seq; (3) two methods for scATAC-seq and one baseline approach. Moreover, we collected four scATAC-seq datasets where the true labels are available. We evaluated the imputation results using three different metrics, including the AUPR-based approach as described above. Our evaluation represents the most comprehensive comparison for scATAC-seq data imputation methods.
- **Biological validation with novel scATAC-seq data:** We successfully applied our approach to novel scATAC-seq data generated from intact mouse kidneys to study fibrosis at different time points. Our analysis identified Runx1 as a critical regulator for myofibroblast differentiation which were validated by biological experiments.



**Figure 1.3: A novel computation method for scATAC-seq data imputation and dimensionality reduction.** This figure shows the workflow of proposed method for scATAC-seq data imputation and dimensionality reduction. An input count matrix first created by counting number of observed reads in each peak and cells. Next, the matrix is normalized by TF-IDF transformation and then factorized into two low-dimensional matrices using NMF. The multiplication of these two matrices represent an imputed and denoised matrix.

## 1.3 Thesis Structure

---

In Chapter 2, we introduce all the concepts needed for the understanding of this thesis. We describe ATAC-seq as a widely used assay for measuring chromatin accessibility at bulk and single-cell resolution. Moreover, we present a computational workflow for bulk and single-cell ATAC-seq data analysis. Finally, we point out the computational challenges for the analysis of scATAC-seq data and review published approaches comprehensively.

In Chapter 3, we present the proposed computational approach to perform data imputation and dimensionality reduction for scATAC-seq data. Our method is based on regularized non-negative matrix factorization (NMF) and term frequency-inverse document frequency (TF-IDF) transformation. Moreover, we describe a method to automatically determine the number of components for NMF.

In Chapter 4, we describe the experiments performed in this thesis to technically and biologically validate our method. We first introduce a novel approach to simulate scATAC-seq, which is mainly used to test the hyper-parameters in our model. Next, we introduce the benchmarking data which is composed of several real-world scATAC-seq datasets. We also describe the computational strategies to benchmark the performance of our approach from different perspectives. Moreover, we describe the experiments performed to generate in-house scATAC-seq and RNA-seq data for biological validation.

In Chapter 5, we outline all analysis results based on our experiments described in Chapter 4. Finally, we close this thesis in Chapter 6 by discussing all presented results, highlighting all the key findings, and pointing out future research opportunities.





### Background

---

In this chapter, we provide the background information required to understand this thesis from both a biological and computational point of view. First, we introduce necessary concepts from molecular biology (Section 2.1), including DNA organization, chromatin accessibility, and gene regulation. Next, we describe biochemical techniques to assess genome-wide chromatin accessibility on bulk and single-cell resolution (Section 2.2). In particular, we will focus on ATAC-seq, a sensitive and straightforward protocol for this task. We then present the workflow for analyzing single-cell open chromatin data (Section 2.3). Afterwards, we give a comprehensive review of competing methods related to this thesis (Section 2.4). Finally, we close this chapter with concluding remarks on the definitions made in this chapter and a brief description of our goal to develop a computational method to analyze single-cell open chromatin data (Section 2.5).

### 2.1 DNA Organization, Chromatin Accessibility, and Gene Regulation

---

We start this section by introducing DNA organization in eukaryotic cells (Section 2.1.1). Next, we describe chromatin accessibility and its function in regulating gene expression. The contents presented in this section are primarily based on Alberts et al. (2017), Pierce (2012), and Klemm et al. (2019).

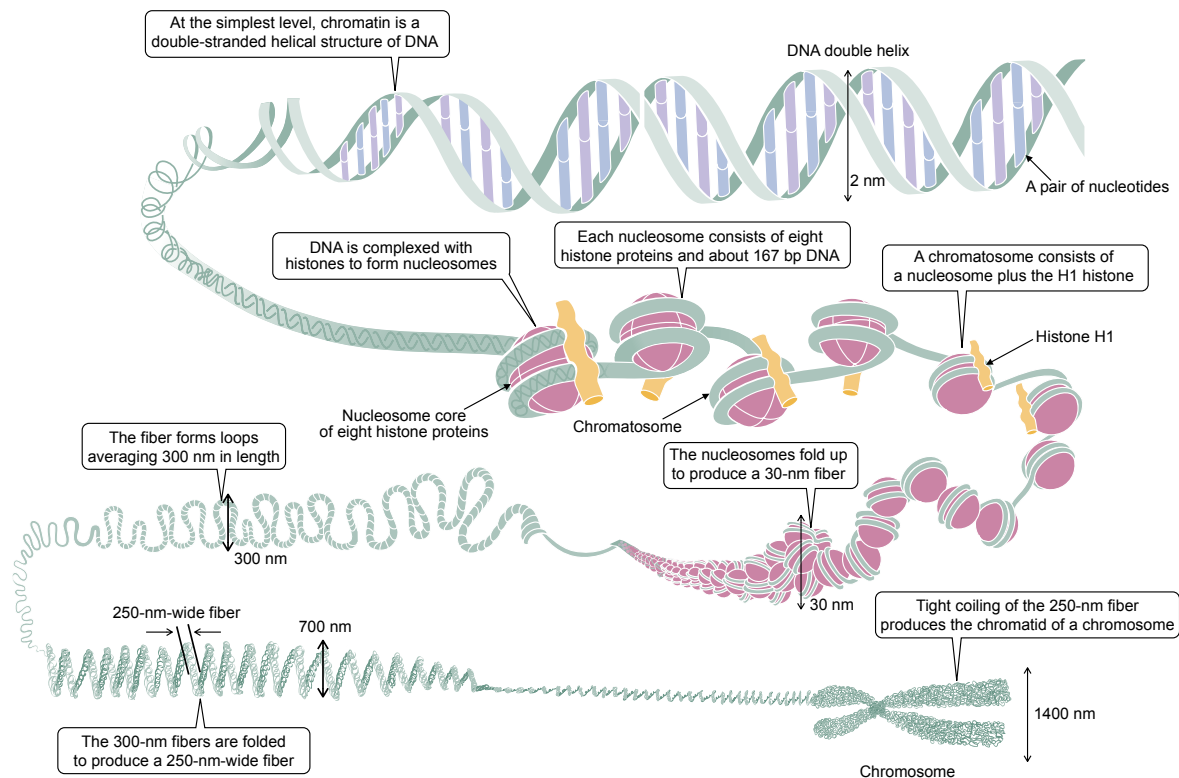
#### 2.1.1 DNA Organization

DNA is macromolecular and carries all genetic instructions for the cell's development, functioning, and reproduction. Each DNA molecule consists of two long polynucleotide chains, known as DNA double strands. Each strand is composed of four types of nucleotide (i.e., adenine (A), cytosine (C), guanine (G), and thymine (T)). The DNA is highly packaged into the chromosome in all eukaryotic organisms. For example, the human chromosome 22 has 48 million nucleotide pairs that would extend for about 1,5 cm if stretched out end-to-end, but it only measures 2  $\mu$ m in length in mitosis, representing a compaction ratio of over 7000-fold (Alberts et al., 2017).

The proteins that bind to DNA in the cell nucleus to help condense it into the chromosome by providing energy are known as histones. The resulting DNA-histone complex, discovered in 1974, is called nucleosome which folds to form chromatin as the basic repeating structural and functional units. In condensed chromatin, adjacent nucleosomes fold on themselves to form a dense and tightly-packed structure that makes up fibers with a diameter of about 30 nm. The next level of chromatin structure is a series of loops of the fibers. Each loop is folded to produce a 250 nm wide fiber, which

## 2.1. DNA Organization, Chromatin Accessibility, and Gene Regulation

produces the chromatid of a chromosome. Figure 2.1 illustrates the organization of DNA from the simple DNA double-helix structure to a complex chromosome.

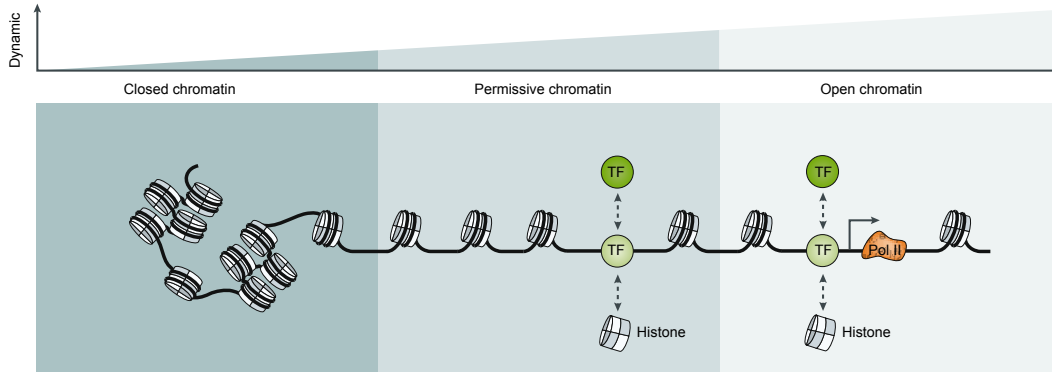


**Figure 2.1: Chromatin organization.** Chromatin has a highly complex structure with several levels of the organization, from simple DNA double helix structure to complex chromosome. Histone proteins bind to DNA to form a nucleosome that folds up to produce a 30 nm fiber. A series of fibers form a loop about 300 nm long, and these 300 nm fibers are further folded to produce a 250 nm wide fiber that is tightly coiled to create the chromatid of a chromosome. *Source: Pierce (2012)* (modified to fit thesis format and/or clarify key points)

### 2.1.2 Chromatin Accessibility and Gene Regulation

Although DNA is tightly compacted into an array of nucleosomes in the eukaryotic cell nucleus, as described above, it nevertheless remains accessible to many enzymes that are responsible for replicating and repairing DNA, and expressing genes in the cell. This physical access to chromatinized DNA is known as chromatin accessibility, a highly dynamic property of chromatin that plays an essential role in establishing and maintaining the cell identity (Klemm et al., 2019). Chromatin accessibility is mainly determined by the occupancy of nucleosomes and other chromatin-binding proteins, such as transcription factors (TFs) (Figure 2.2).

Gene regulation is the process by which the cell determines which genes will be active ("ON") and which genes will be inactive ("OFF"). It is the key for cell specification to perform particular functions in multicellular eukaryotic organisms, and is substantially determined by chromatin acces-



**Figure 2.2: Chromatin accessibility dynamics.** A continuum of accessibility status reflects the distributions of chromatin dynamics across the genome. In contrast to closed chromatin, permissive chromatin is sufficiently dynamic for transcription factors to initiate sequence-specific accessibility remodeling and establish an open chromatin conformation. TF, transcription factor; Pol II, polymerase II. *Source: Klemm et al. (2019)* (modified to fit thesis format and/or clarify key points)

sibility. This means that in different biological systems, for instance, different cell types or the same cell type under distinct conditions (e.g., health vs. disease), the same genomic loci can have different chromatin states (i.e., open versus closed). These genomic regions are known as *cis*-regulatory elements (CREs) and are associated with gene regulation. More specifically, the gene transcription rates are controlled by TFs which bind to a specific set of DNA sequences, called transcription factor binding sites (TFBSs). The binding preference for a particular TF can be modeled with a position weight matrix (PWM) and visualized as a sequence logo (Figure 2.3).

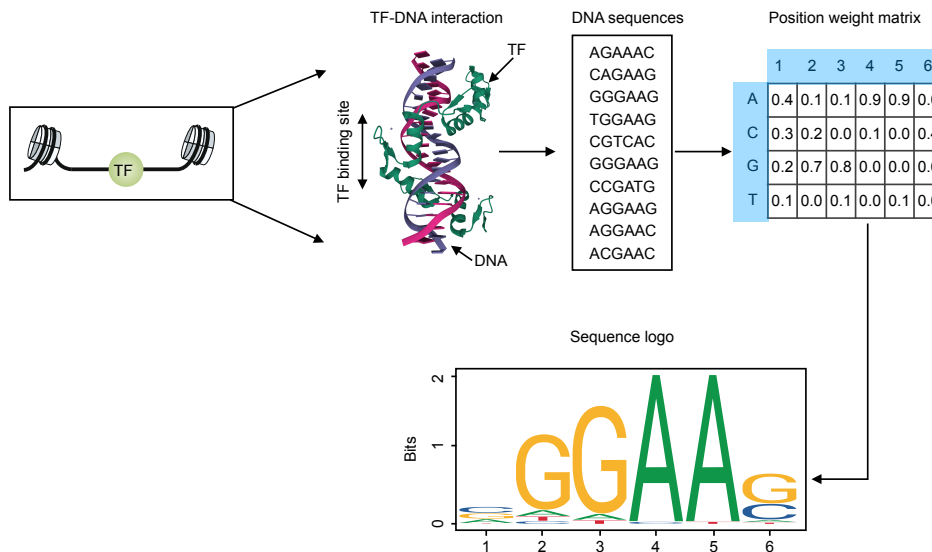
## 2.2 Profiling Chromatin Accessibility with ATAC-seq

Over the last decade, a wide variety of biochemical methods have been developed to quantitatively measure chromatin accessibility in a genome-wide manner using next-generation sequencing (NGS) technology. Among these methods, ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) (Buenrostro et al., 2013, 2015a) has recently become the standard protocol because it has a low requirement for the number of input cells and needs less preparation time for library construction. Moreover, it has been further improved to accommodate single-cell sequencing to assess chromatin accessibility at single-cell resolution. In this section, we first introduce bulk ATAC-seq (Section 2.2.1). Next, we describe single-cell ATAC-seq (scATAC-seq) in Section 2.2.2. In particular, we will focus on droplet-based scATAC-seq, as it has been widely used by researchers and is commercially available by 10X Genomics.

### 2.2.1 Bulk ATAC-seq

ATAC-seq was introduced in 2013 as an advanced method to profile open chromatin (Buenrostro et al., 2013). In ATAC-seq, the genetically engineered hyperactive Tn5 transposase is loaded *in vitro* with two adapters for high-throughput DNA sequencing. Therefore, it can simultaneously fragment

## 2.2. Profiling Chromatin Accessibility with ATAC-seq



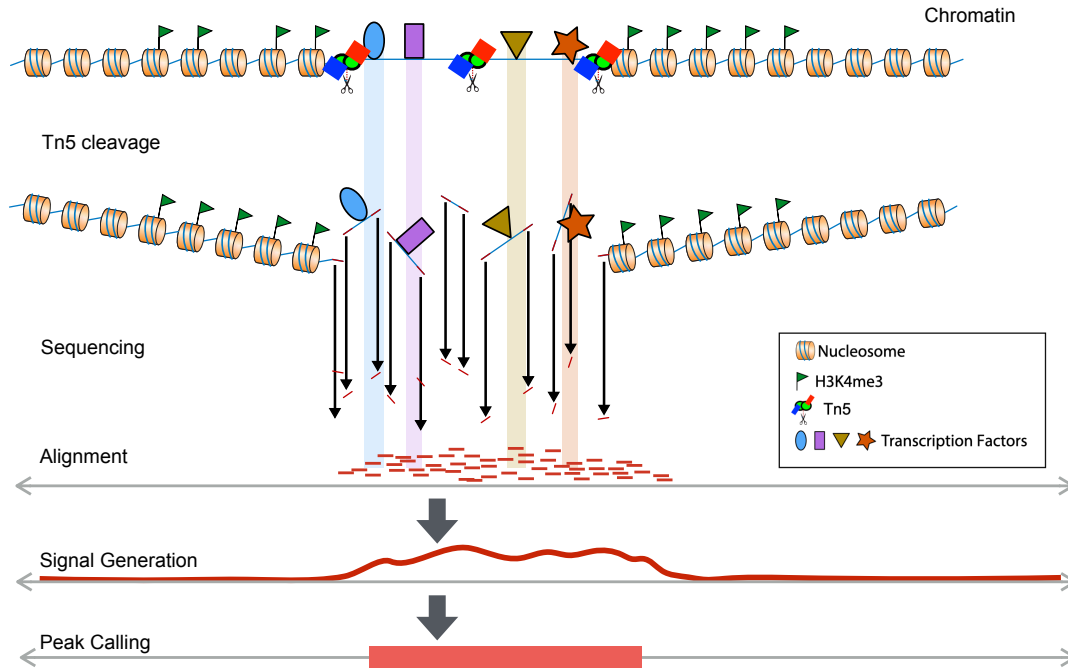
**Figure 2.3: Specificity of TF-DNA interaction.** TF binds to specific DNA sequences. These sequences can be represented using a position weight matrix where each row represents a nucleotide, and each column represents a position. The value denotes the probability of a nucleotide at a particular position. This matrix can also be visualized as a sequence logo. *Source: Klemm et al. (2019)* (modified to fit thesis format and/or clarify key points).

and tag genomic regions from accessible chromatin. The low-input requirement of ATAC-seq makes it possible to profile chromatin accessibility for biological processes consisting of a very limited number of cells, such as preimplantation development (PD) and zygotic genome activation (ZGA) of early zygotes (Bentsen et al., 2020). Given the sensitivity and simplicity of the protocol, ATAC-seq is currently the most commonly used method for measuring chromatin accessibility in research laboratories. For example, the number of studies using ATAC-seq protocol deposited in Gene Expression Omnibus (GEO) is about five times higher than the number of studies using DNase-seq protocol (16,468 ATAC-seq versus 3,454 DNase-seq), as queried using the words "ATAC-seq" and "DNase-seq" on May 22, 2021.

ATAC-seq begins with cell collection which typically contains 500-50,000 cells. After transposition, the transposed DNA fragments are PCR (polymerase chain reaction) amplified and sequenced. In this step, paired-end sequencing is usually preferred as it provides information about fragment size which can be used to infer nucleosome position (Schep et al., 2015) and improve transcription factor binding sites prediction (Li et al., 2019). The sequenced DNA fragments are then mapped to the reference genome to obtain aligned DNA reads, and an ATAC-seq signal can be created by counting the number of cutting events per genomic position. Finally, the open chromatin regions are identified by performing peak calling (Figure 2.4).

### 2.2.2 Single Cell ATAC-seq

Although ATAC-seq is able to profile chromatin accessibility from low-input samples, the obtained profiles are still averaged across all input cells, thus masking heterogeneity between and within cell types. To address this issue, single-cell ATAC-seq (scATAC-seq) was introduced in 2015 (Buenrostro

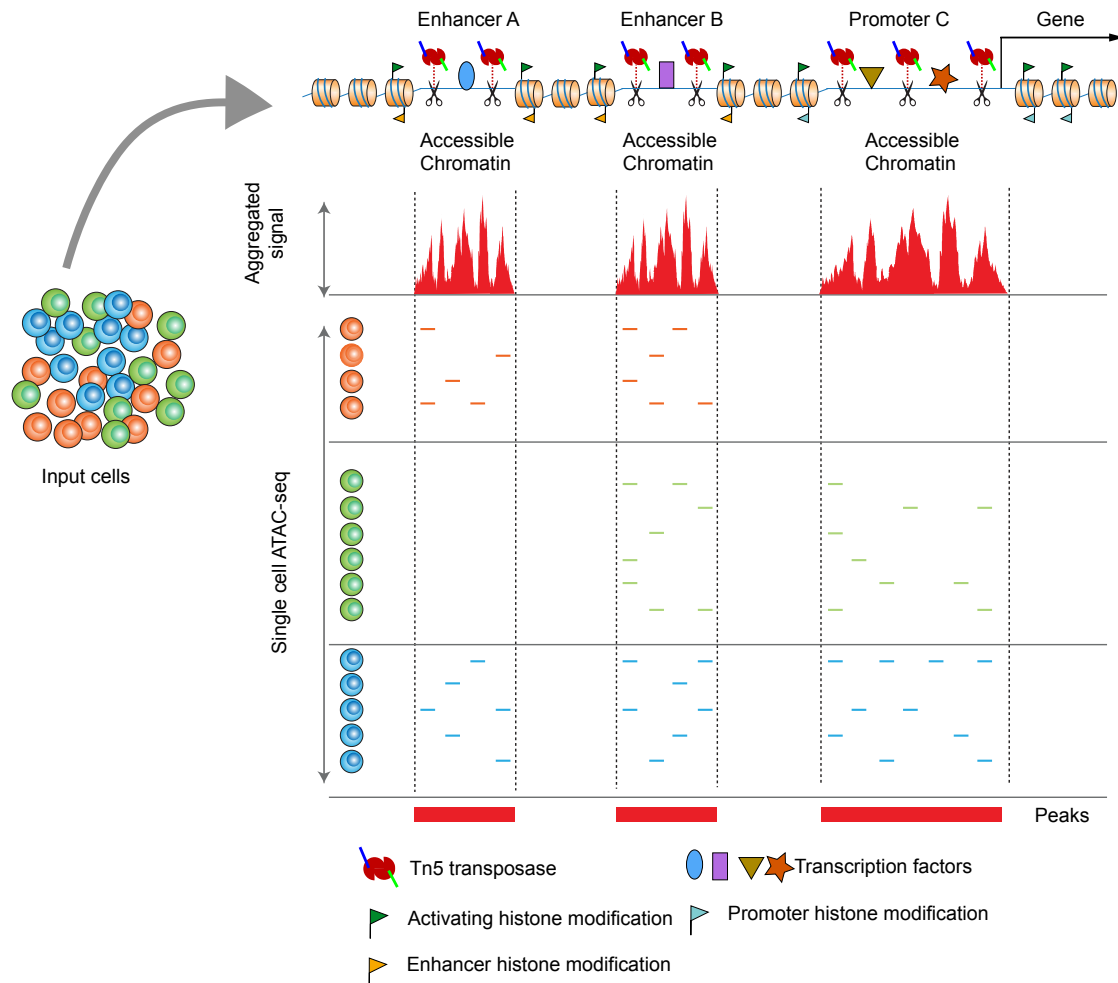


**Figure 2.4: Measuring chromatin accessibility using ATAC-seq.** This figure shows the overall workflow of ATAC-seq to profile chromatin accessibility. The Tn5 enzyme is used to cleave and tag double-stranded DNA with sequencing adapters. Next, the DNA fragments are sequenced and mapped to a reference genome. Then, a genome-wide signal is generated by counting the number of aligned reads per genomic coordinate. Finally, chromatin accessible regions are detected with peak calling algorithm.

et al., 2015b; Cusanovich et al., 2015). With computational approaches, scATAC-seq enables the identification of cell-type-specific *cis*-regulatory elements (CREs) at single-cell resolution and the recovery of novel cell types from the input samples (Figure 2.5).

There are several techniques to implement scATAC-seq. One of the most used approaches is called droplet-based scATAC-seq, in which the cells are captured through a microfluidic device. More specifically, nuclei are first isolated from a single-cell suspension and transposed in bulk with transposase Tn5. Next, transposed nuclei are loaded onto a microfluidic chip for the generation of gel beads in emulsion (GEM). Each gel bead is barcoded with single-stranded oligonucleotides that consist of a 29-bp sequencing adapter, a 16-bp barcode selected from 750,000 designed sequences to index GEMs, and the first 14 bp of read 1N, which serves as the priming sequence in the linear amplification reaction to incorporate barcodes to transposed DNA. After GEM generation, gel beads are dissolved, and the oligonucleotides are released for linear amplification of transposed DNA. Finally, the droplet emulsion is broken, and barcoded DNA fragments are pooled for PCR amplification to generate indexed libraries for high-throughput sequencing. This technique can profile accessible chromatin from tens of thousands of cells with high data quality in each experiment. Also, it has been commercialized by 10x Genomics (Chromium Next Gem Single Cell ATAC-seq Library Kit) (Satpathy et al., 2019), making it the most commonly used protocol for profiling single-cell chromatin accessibility (Satpathy et al., 2019) (Figure 2.6).

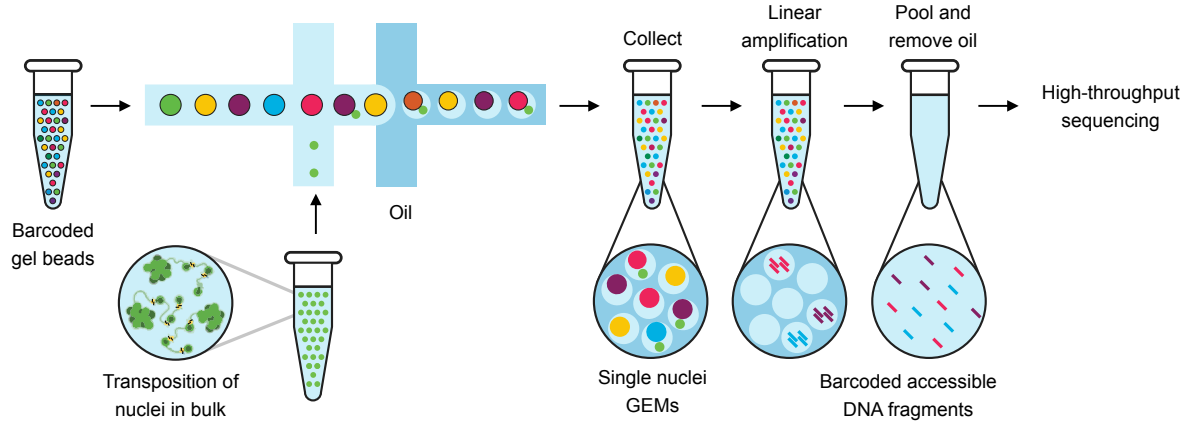
## 2.3. Computational Analysis of ATAC-seq



**Figure 2.5: Measuring chromatin accessibility using single cell ATAC-seq.** This figure shows the overall experimental principle of ATAC-seq to profile chromatin accessibility from a heterogeneous sample containing three different cell populations at single-cell resolution. The output from single-cell ATAC-seq is a sparse matrix where each row represents a cell and each column presents a peak.

## 2.3 Computational Analysis of ATAC-seq

In this section, we describe the computational workflow for analysis of bulk (Section 2.3.1) and single-cell ATAC-seq (Section 2.3.2) data, respectively. Because the inputs for both analyses are massive amount of short sequencing reads, they share the same low-level processing procedures, including quality control and alignment, like other high throughput sequence data. The major difference between single-cell and bulk ATAC-seq data is that every read in scATAC-seq library carries a unique cell barcode that is used to distinguish different cells. Because of that, the analysis of bulk ATAC-seq is usually performed based on all reads from the library, while analyzing scATAC-seq data requires additional steps, such as cell calling and matrix creation, to first recovery and separate cells. The following analysis for scATAC-seq usually takes the count matrix as input.



**Figure 2.6: Droplet-based scATAC-seq.** Schematic of droplet-based scATAC-seq. Nuclei are isolated and transposed with Tn5 transposase. Next, GEM is generated with barcodes and subjected to linear amplification. Finally, the emulsion is broken and barcoded DNA fragments are pooled and sequenced. *Source: Satpathy et al. (2019)* (modified to fit the thesis format and/or clarify key points)

### 2.3.1 Bulk ATAC-seq

Analysis of bulk ATAC-seq data generally consists of three main components: alignment, peak calling, and computational footprinting analysis. We consider the alignment as data preprocessing, peak calling as core analysis, and computational footprinting as downstream analysis for bulk ATAC-seq.

#### Alignment

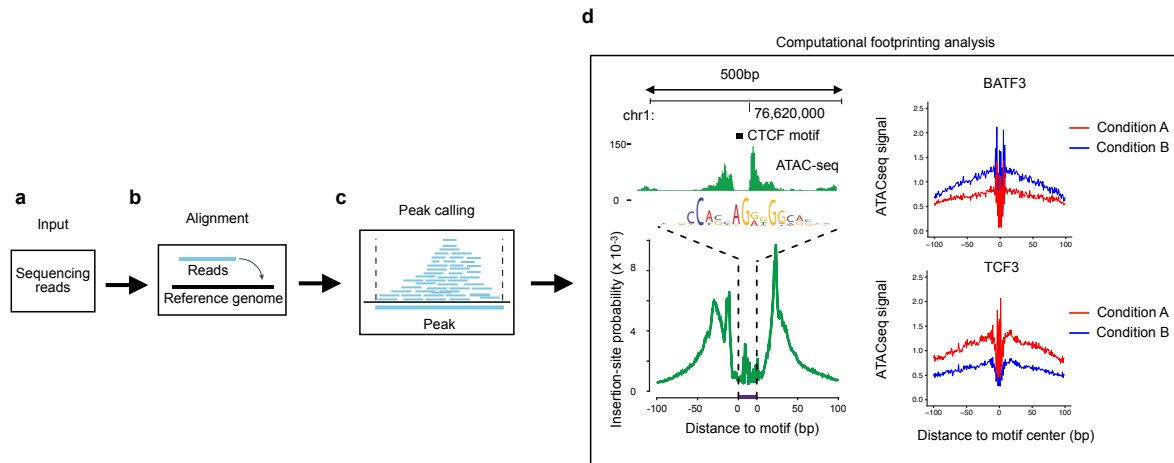
This is the very first step for computational analysis of ATAC-seq data. It takes the raw sequencing reads as input and finally outputs an alignment file containing all aligned DNA reads by mapping the reads to a reference genome (Figure 2.7a-b). More formally, this problem can be defined as: given a large reference genome (usually up to billions of base pairs) and millions of short reads (<200 bps), find the most probable position of the reads in the genome. There are several tools available to perform the alignment, of which Bowtie2 (Langmead and Salzberg, 2012) and BWA (Li and Durbin, 2009) are the most popular ones. The former is used in the ENCODE ATAC-seq pipeline, and the latter is used in the Cell Ranger ATAC pipeline by 10X Genomics.

#### Peak Calling

Peak calling plays an essential role in chromatin accessibility data analysis (Figure 2.7c). It aims to detect the genomic regions with a high accumulation of reads compared with the expectation (i.e., background). These regions are defined as accessible chromatin regions (also referred to as peaks). More formally, the peak calling problem is defined as find genomic regions of arbitrary size with more reads than expected by chance. A simply peak caller can be implemented in two steps: 1) using a fix window to scan through the genome to obtain a distribution of counts per bin, 2) defining a statistical test to evaluate if the number of reads is higher than expected by chance. However, in



## 2.3. Computational Analysis of ATAC-seq



**Figure 2.7: Computational analysis of bulk ATAC-seq.** **a**, The input for ATAC-seq data analysis is the raw sequencing reads. **b**, Next, the sequencing reads are aligned to a reference genome by finding the most probable positions. **c**, Based on the aligned reads, peak calling is performed to identify the accessible regions, which are usually used as features in downstream analysis. **d**, For data interpretation, computation footprinting method (e.g., HINT-ATAC) is used to identify transcription factor binding sites for ATAC-seq. Moreover, the differential analysis of TF footprints can be used to compare TF activity between different conditions. *Source: Buenrostro et al. (2013)* (modified to fit thesis format and/or clarify key points).

reality, proper quantification of read counts require several future steps, such as CG bias correction, duplicated reads, mappability, and fragment size estimation. Currently, the most widely used peak calling tool for ATAC-seq is MACS2 (Zhang et al., 2008), which was initially designed for ChIP-seq data. On the other hand, HMMRATAC (the Hidden Markov ModelER for ATAC-seq) is the first dedicated tool to identify peaks for ATAC-seq by considering the Tn5 digested DNA fragments that contain additional nucleosome positioning information (Tarbell and Liu, 2019).

### Computational Footprinting Analysis

In addition to profile genome-wide open chromatin states, ATAC-seq can also be used to identify nucleotide-level transcription factor binding sites (TFBSs) via computational search footprint-like regions with low numbers of Tn5 cuts surrounded by regions with high numbers of cuts (Buenrostro et al., 2013). This process is called computational footprinting analysis (Gusmao et al., 2016). The intuition is that the binding of a TF will protect DNA from Tn5 cleavage, thus creating a footprint-like shape in nucleotide resolution ATAC-seq signal (Figure 2.7d). By detecting genomic regions with such a pattern, active TFBSs can be uncovered. However, TF footprinting for ATAC-seq has some limitations as it requires a deeply sequenced library and is biased by intrinsic sequence preference for the Tn5 enzyme. To address these issues, we have developed HINT-ATAC, the first computational footprinting method tailored to the ATAC-seq protocol, by correcting Tn5 cleavage bias and considering strand-specific bias of ATAC-seq signal (Li et al., 2019). Moreover, HINT-ATAC also allows for comparing the TF activity between different conditions by using differential footprinting analysis.



### 2.3.2 Single Cell ATAC-seq

We here describe the computational workflow for scATAC-seq data analysis, including preprocessing (cell calling and matrix construction), core analysis (dimensionality reduction and clustering), and downstream analysis (cell annotation, motif analysis, and co-accessibility analysis).

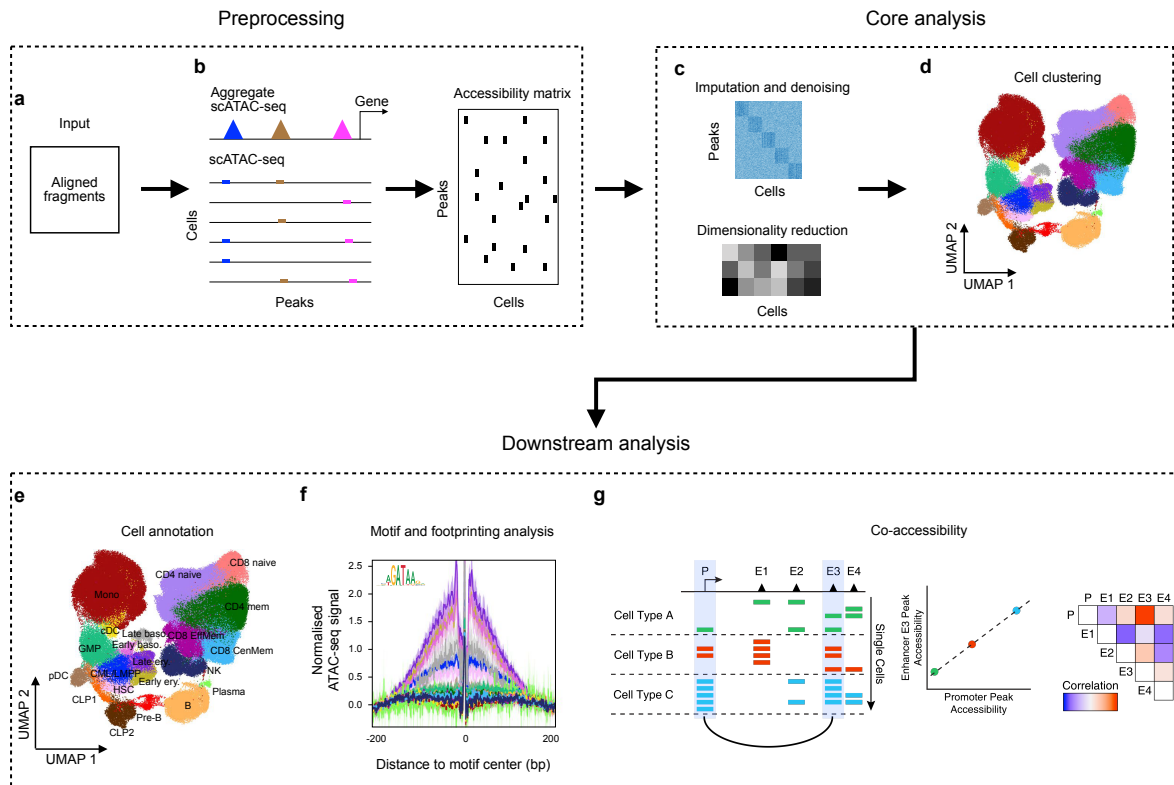
#### Preprocessing

In addition to read alignment, preprocessing of scATAC-seq requires two additional steps: cell calling and matrix construction. The aim is to generate a peak by cell count matrix for the following analysis.

**Cell calling** This is a unique and essential step when dealing with scATAC-seq data. In this step, barcodes with high-quality, aligned reads are identified as valid cells, and all other non-cell barcodes will be excluded from the data. The barcodes are a set of unique and short DNA sequences (usually 16 bases) that are carried by gel beads to distinguish different cells (Figure 2.6). The metrics used for calling cells include the number of unique fragments and signal-to-noise ratio (SNR). The first metric is straightforward, as cells with few available fragments usually cannot provide enough information for interpretation and should be removed. The commonly used threshold in the community is 1,000. For the second metric, there are several ways to calculate SNR using scATAC-seq. One of them is known as FRiP (fraction of reads in peaks). It is worth noting that to calculate FRiP, a pre-defined peak set obtained by performing peak calling using all cells is needed, which might not capture enough signal from rare cells, leaving these cells undetected in clustering analysis. Another option is to use the TSS enrichment score, which is defined as the ratio between the aggregate distribution of reads that centered on TSSs and those flanking the corresponding TSSs. This metric has been widely used to measure the quality of bulk ATAC-seq data in the ENCODE project. The intuition is that ATAC-seq data is universally enriched at gene TSS regions, compared to other genomic regions, due to large protein complexes that bind to the promoter. Another approach is to use fragments in promoter ratio, similar to the TSS enrichment score but with a lower computational complexity.

**Matrix Construction** To analyze the scATAC-seq, a binary accessibility matrix is created (Figure 2.8b). scATAC-seq has a different readout comparing with scRNA-seq (i.e., chromatin accessibility versus gene expression). For scRNA-seq, one can directly use genes as features to build a count matrix in which each element represents the measured expression of a gene in a particular cell. However, in scATAC-seq, the features are unavailable and need to be defined depending on the downstream analysis. The most common selection is to use the accessible regions or peaks obtained by performing peak calling on the bulk ATAC-seq or aggregate scATAC-seq data. Notably, the peaks sometimes are dominated by the major cell types present in the data, and signals from rare cells can be neglected. Alternatively, the features can be defined as equal-size window along the genome. One challenge of using this approach is how to choose the optimal bin size and the large number of the genome.

## 2.3. Computational Analysis of ATAC-seq



**Figure 2.8: Analytic workflow of scATAC-seq.** **a**, Preprocessing of scATAC-seq data starts with cell-specific aligned fragments. **b**, Next, accessible regions are identified based on aggregate scATAC-seq, and an accessibility matrix is created by counting the number of reads per cell per peak. **c**, Based on the accessibility matrix, imputation and denoising is applied to obtain an imputed matrix, and dimensionality reduction is performed to learn a low-dimensional representation for the cells. Cell clustering can be done with both imputation matrix or dimensionally reduced matrix. **d**, Unsupervised clustering analysis is used to group cells. The colors refer to clusters. **e**, Cell types are identified by annotating the obtained clusters, as represented by different colors. **f**, Line plots visualizing motif footprint profiles generated using aggregate ATAC-seq data across cell types for TF GATA. **g**, Illustration of co-accessibility analysis. Here, enhancer E3 is often co-accessible with promoter P across cell type A, B, and C, suggesting a potential regulatory interaction between enhancer E3 and promoter P. *Source: Granja et al. (2021)* (modified to fit thesis format and/or clarify key points).

### Core Analysis

The core analysis of scATAC-seq involves dimensionality reduction of the scATAC-seq matrix and cell clustering. The goal is to assign a cluster label to each of the cells.

**Dimensionality Reduction** The inherent high dimensionality in scATAC-seq data makes it difficult for further analysis (e.g., cell clustering and visualization). Therefore, it is essential to reduce the dimensions of the accessibility matrix by projecting the data into a low-dimensional space while preserving similarity between the cells (Figure 2.8c). One of the most commonly used methods is PCA (Principal Component Analysis), which linearly projects the high-dimension

data onto only the first few principal components to obtain a low-dimensional representation while retaining as much data variation as possible. However, it is inappropriate to apply the same strategy to scATAC-seq due to the low count issue and sparsity of the data. For example, in a single cell, a chromatin region can either be open or closed, which produces a binary matrix where 1 represents an accessible region in a cell and 0 is inaccessible otherwise. Moreover, even in high-quality scATAC-seq data, the majority of accessible regions are not transposed, causing most of the entries in the matrix to be 0. To address these problems, researchers have either adapted existing methods from other fields, such as LSI (latent semantic indexing) from natural language processing, or developed novel techniques (e.g., SnapATAC). We will detail these methods in Section 2.4.3

**Cell Clustering** Unsupervised clustering is of central importance for single-cell data analysis. It is used to identify putative cell types and substantially impacts the biological interpretation of the data (Figure 2.8d). Many algorithms are available, and they have significant differences in considering what constitutes a cluster and how to find them efficiently. Currently, most single-cell methods focus on computing nearest neighbor graphs in reduced dimensions and then detect "communities", in which cells are densely connected. The most popular algorithm for this task is called Louvain (Blondel et al., 2008), which has been found to work exceptionally well and is now a standard practice in the single-cell field (Stuart et al., 2019, 2020; Granja et al., 2021), although many others are also available (Xie et al., 2013). Recently, a new method, named scABC (single-cell Accessibility Based Clustering), was developed to perform unsupervised clustering specifically for scATAC-seq data, but it is not widely used in the community. In sum, for scATAC-seq clustering, the community-detection algorithms are frequently used, given their speed and accuracy.

## Downstream Analysis

After clustering the cells, downstream analysis is needed to characterize the molecular mechanisms and understand the biological system. Usually, it consists of cell annotation, motif analysis, and co-accessibility analysis.

**Cell Annotation** In this step, the goal is to identify the cell types that these clusters represent (Figure 2.8e). For scATAC-seq, this is very challenging because less is known about the functional role of non-coding genomic regions. In order to use marker genes to annotate cells, as commonly used in the scRNA-seq analysis, one can infer gene activity scores from scATAC-seq. The intuition is that if a gene is expressed in a cell, then the chromatin around this gene must be accessible. The simplest model is to quantify the chromatin accessibility associated with each gene by summing up the fragments intersecting the gene body and promoter region (Stuart et al., 2020).

**Motif and Footprinting Analysis** DNA sequence motif analysis provides an alternative way to understand the scATAC-seq data. In this analysis, the goal is to identify TFs that are relevant for

## 2.4. Related Works

different cell types. For this, one can use differential footprinting analysis (Li et al., 2019; Stuart et al., 2020; Granja et al., 2021). It first detects the TF footprints based on the pseudo-bulk data for each cluster or cell type and then quantifies the differences between the footprint profiles (Figure 2.8f). Of note, this approach usually requires a relatively deep sequencing library which in some cases (e.g., rare cells) is not possible to obtain. Nevertheless, it can provide a clear visualization of TF binding sites accessibility across different conditions.

**Co-accessibility analysis** Co-correlation of accessible peaks across individual cells has been reported to have a high agreement with previously observed chromosome compartments (Buenrostro et al., 2015b; Kalhor et al., 2012). Based on this observation, an algorithm, called Cicero, was developed to identify co-accessible pairs of *cis*-regulatory elements (Pliner et al., 2018). The resulting predictions are called peak-to-peak links and can provide an idea of the overall structure of the *cis*-architecture of a genomic region. Furthermore, peak-to-gene links can also be obtained if the peaks are linked to genes by estimating the correlation of the chromatin accessibility between the peaks and promoters. In doing so, one can identify the putative target genes for each peak and build a gene regulatory network in a data-driven manner. Additionally, comparison of the predicted links between different conditions, such as health vs. disease, allows the characterization of chromatin structure dynamics. This can advance our quantitative understanding of eukaryotic gene regulation (Figure 2.8g).

**Trajectory Analysis** Trajectory analysis, also known as pseudotime analysis, aims to computationally order the cells based on the progression through a developmental process using single-cell sequencing data. This concept was first introduced to analyze scRNA-seq data (Trapnell et al., 2014). Since then, more than 70 trajectory inference tools have been developed. The most widely used tool for trajectory inference is called Monocle 3 (Cao et al., 2019), which has been implemented by Signac to build trajectory for scATAC-seq data. On the other hand, ArchR first manually creates a trajectory backbone in the form of an ordered vector of cell groups labels and then fits a supervised trajectory in a low-dimensional space (Granja et al., 2021). The evaluation shows that both approaches can recover known cell development processes for hematopoietic cells (Granja et al., 2021). In this thesis, we will use ArchR to perform trajectory analysis.

## 2.4 Related Works

---

In this section, we present the state-of-the-art computational methods for scATAC-seq data analysis. In particular, we will focus on the algorithms for matrix imputation and dimensionality reduction, as they serve as the core analysis for scATAC-seq. More specifically, we first point out some challenges when analyzing scATAC-seq, which motivates this thesis (Section 2.4.1). Next, We give a comprehensive literature review on published computational methods for single-cell data imputation (Section 2.4.2) and dimensionality reduction (Section 2.4.3).

### 2.4.1 Challenges of Analysis scATAC-seq

The analysis of scATAC-seq data is a newer area compared with scRNA-seq, for which tools such as Seurat (Stuart et al., 2019) and Scanpy (Wolf et al., 2018) have been developed for years and considered as a standard practice. Although they share many properties, it is more challenging to analyze scATAC-seq data than scRNA-seq data for three reasons:

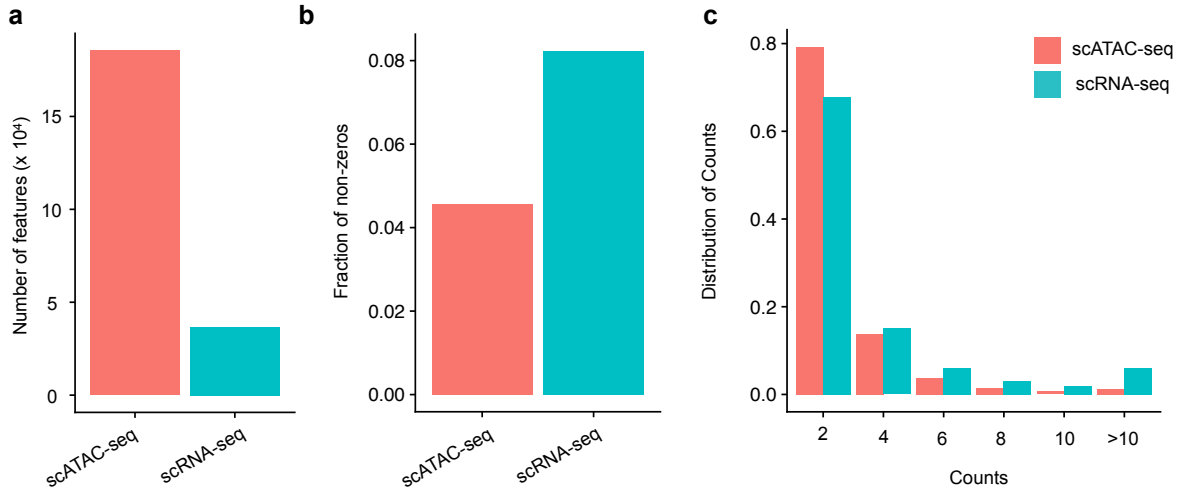
- Single-cell ATAC-seq data has much more features than scRNA-seq because multiple *cis*-regulatory modules often regulate a gene (De-Leon and Davidson, 2007). In fact, the number of features in scATAC-seq is usually one order magnitude higher than that in scRNA-seq (usually  $10^5$  regions vs.  $10^4$  genes), as shown in Figure 2.9a.
- Single-cell ATAC-seq data is sparser than scRNA-seq. As with other single-cell data, scATAC-seq is also affected by dropout events due to the loss of DNA material during library preparation. Moreover, the number of DNA molecular copies is limited comparing with message RNA (mRNA). Together, these characteristics render the scATAC-seq count matrix extremely sparse. In contrast, scRNA-seq has less severe sparsity than scATAC-seq due to smaller dimensions and lower dropout rates for genes with high or moderate expression levels (Figure 2.9b).
- Single-cell ATAC-seq data has a lower count than scRNA-seq (Figure 2.9c). A particular chromatin region can either be accessible on one allele, both alleles, or no alleles in a single cell. Therefore, only two Tn5 insertion events are expected for a region per cell. For example, an observation of two Tn5 cleavage fragments from a single region in a single cell cannot confidently determine that this region in this cell is two times more accessible than another cell that only has one Tn5 insertion. For this reason, it makes more sense to work on a binarized scATAC-seq data matrix, where 1 means accessible and 0 means non-accessible.

### 2.4.2 Computational Methods for Single Cell Data Denoising and Imputation

The single-cell sequencing data is arguably sparse and noisy because of the low amounts of starting material from each cell and stochastic fluctuations of cell states (Kharchenko, 2021). Consequently, the observed non-zeros may not coincide with the true abundance in a single cell. Moreover, the observed zero values may be either due to truly non-expressed (for scRNA-seq) or non-accessible (for scATAC-seq) or technical limitations of the sequencing technology (Patrino et al., 2021).

Many computational methods have been developed to address these issues, and they can be broadly categorized into two groups: (i) denoising methods and (ii) imputation methods. Although both approaches take as input the raw count matrix and output an estimated matrix with the exact dimensions, they are based on different assumptions and employ different computational strategies to accomplish their tasks. Specifically, denoising methods assume that the data has some technical noise and try to remove the noise by adjusting the values. In contrast, imputation methods assume that the existing values (i.e., non-zero) are correct and aim to recover the missing entries (zeros) in the data. However, these two terms are often used interchangeably in the field.

## 2.4. Related Works



**Figure 2.9: Statistical comparison of scATAC-seq and scRNA-seq data.** **a**, Barplot comparing the number of features between scATAC-seq and scRNA-seq. **b**, Barplot comparing the fraction of non-zeros between scATAC-seq and scRNA-seq. **c**, Barplot comparing the distribution of counts in the matrices of scATAC-seq and scRNA. The data was generated using the 10x Multiome protocol, which simultaneously profiles gene expression and open chromatin from the same cell. Therefore, the number of cells in scATAC-seq and scRNA-seq are the same, which allows for a fair comparison across modalities.

We here review the published computational methods for single-cell data denoising and imputation. It is worth pointing out that comparing to scRNA-seq, denoising and imputation for scATAC-seq data remains significantly unexplored, and only a few algorithms are tailored for this task. Therefore, approaches presented here are primarily developed for scRNA-seq and are adapted into scATAC-seq in this thesis for evaluation. Moreover, these approaches can be categorized into four groups based on the techniques that they used. The first group contains data-smoothing methods that define a similarity between cells and then adjust expression values for each cell based on the values in similar cells (MAGIC). The second group includes model-based methods that model the observed values in each cell as a random variable and perform imputation (scImpute) and denoising (SAVER). The third group is composed of methods that use a deep learning approach to learn a latent space representation of the cells and then denoise the data by reconstructing the input matrix (DCA; SCALE). Finally, the last group includes methods that denoise the data by solving a matrix factorization problem (scBFA). Table 2.1 summarizes the computational methods, and Figure 2.10 shows the workflow of each method.

### MAGIC

MAGIC (Markov Affinity-based Graph Imputation of Cells) is, to the best of our knowledge, the first method for explicitly imputing scRNA-seq data (Van Dijk et al., 2018). It learns the manifold of the data and uses the resulting graph to smooth the features and restore the structure of the data. MAGIC takes an observed count matrix as input and produces an imputed matrix representing the likely gene expression of individual cells. It first constructs an affinity matrix using an adaptive Gaussian kernel

from which a Markov transition matrix is created, which represents the probability distribution of transitioning from one cell into another in a single step. Next, MAGIC raises this matrix to the power  $t$  to produce a matrix where each entry represents the probability that a random walk of length  $t$  starting at cell  $i$  will reach cell  $j$ , a process known as data diffusion. MAGIC multiplies the transition matrix by the original data matrix for imputation, restoring cells to the underlying manifold. Figure 2.10a depicts the workflow of the MAGIC algorithm.

### scImpute

scImpute is a model-based imputation method (Li and Li, 2018). It first normalizes the input count matrix by the library size of each cell so that all cells have one million reads. Next, scImpute clusters the cells into  $K$  sub-populations and selects a number of candidate neighbors for each cell from the same cluster. Instead of treating all zero values as dropouts, scImpute infers which genes are affected by dropout events in which cells by constructing a statistical model. Specifically, it uses a mixture model of two components to model the gene expression with dropout events. The first component is a Gamma distribution used to account for the dropout events, and the second component is a Normal distribution to represent the actual gene expression levels. Then for each gene  $i$ , its expression in cell population  $k$  is modeled as a random variable  $X_i^k$  with density function:

$$f_{X_i^k}(x) = \lambda_i \cdot \text{Gamma}(x; \alpha_i^k, \beta_i^k) + (1 - \lambda_i) \cdot \text{Normal}(x; \mu_i^k, \sigma_i^k) \quad (2.1)$$

where  $\lambda_i$  is the dropout rate of gene  $i$  in sub-population  $k$ ,  $\alpha_i^k, \beta_i^k$  are the shape and rate parameters of Gamma distribution, and  $\mu_i^k, \sigma_i^k$  are the mean and standard deviation of the Normal distribution. The parameters are estimated using the Expectation-Maximization (EM) algorithm and are denoted as  $\hat{\lambda}_i, \hat{\alpha}_i^k, \hat{\beta}_i^k, \hat{\mu}_i^k$  and  $\hat{\sigma}_i^k$ . Then, the dropout probability of gene  $i$  in cell  $j$  can be calculated as:

$$d_{ij} = \frac{\hat{\lambda}_i \cdot \text{Gamma}(X_{ij}; \hat{\alpha}_i^k, \hat{\beta}_i^k)}{\hat{\lambda}_i \cdot \text{Gamma}(X_{ij}; \hat{\alpha}_i^k, \hat{\beta}_i^k) + (1 - \hat{\lambda}_i) \cdot \text{Normal}(X_{ij}, \hat{\mu}_i^k, \hat{\sigma}_i^k)}. \quad (2.2)$$

Next, for each cell  $j$ , scImpute selects a gene set  $A_j$  in need of imputation based on a threshold  $t$  on dropout probability. The rest of the genes form a gene set  $B_j$  that does not need imputation. Finally, the expression of genes in  $A_j$  is imputed by borrowing information from similar cells based on gene set  $B_j$ . The imputation workflow of scImpute is shown in Figure 2.10b.

### SAVER

SAVER (Single-cell Analysis Via Expression Recovery) is a Bayesian model aiming to recover the true expression of each gene in each cell by borrowing information across genes and cells (Huang et al., 2018) (Figure 2.10c). It models the observed expression of gene  $g$  in cell  $c$   $Y_{gc}$  using the following function:

$$\begin{aligned} Y_{gc} &\sim \text{Poisson}(s_c \lambda_{gc}) \\ \lambda_{gc} &\sim \text{Gamma}(\alpha_{gc}, \beta_{gc}) \end{aligned} \quad (2.3)$$

## 2.4. Related Works

where  $\lambda_{gc}$  represents the normalized true expression,  $s_c$  represents the size normalization factor,  $\alpha_{gc}, \beta_{gc}$  are the shape and rate parameters of Gamma distribution, a prior of  $\lambda_{gc}$ . The goal of SAVER is to derive the posterior gamma distribution for  $\lambda_{gc}$  given the observed count  $Y_{gc}$  and use the posterior mean as the normalized SAVER estimate  $\hat{\lambda}_{gc}$ . To do so, SAVER estimates the prior mean  $\mu_{gc}$  and variance  $v_{gc}$  using an empirical Bayes-like technique. Specifically, it first uses the log-normalized counts of all other genes  $g'$  in the same cells as predictors and fits a Poisson generalized linear regress model with a log link function:

$$\log(\mu_{gc}) = \gamma_{g0} + \sum_{g' \neq g} \gamma_{gg'} \log\left(\frac{Y_{g'c} + 1}{s_c}\right). \quad (2.4)$$

The prediction  $\hat{\mu}_{gc}$  is treated as the prior mean for each gene in each cell. Next, SAVER estimates the prior variance  $\hat{v}_{gc}$  by assuming an underlying mean-variance function for a given gene  $g$ . Given these parameters, the posterior distribution can be written as:

$$\lambda_{gc} | Y_{gc}, \hat{\alpha}_{gc}, \hat{\beta}_{gc} \sim \text{Gamma}(Y_{gc} + \hat{\alpha}_{gc}, s_c + \hat{\beta}_{gc}). \quad (2.5)$$

Moreover, the posterior mean can be estimated by:

$$\hat{\lambda}_{gc} = \frac{Y_{gc} + \hat{\alpha}_{gc}}{s_c + \hat{\beta}_{gc}} = \frac{s_c}{s_c + \hat{\beta}_{gc}} \cdot \frac{Y_{gc}}{s_c} + \frac{\hat{\beta}_{gc}}{s_c + \hat{\beta}_{gc}} \cdot \hat{\mu}_{gc}. \quad (2.6)$$

## DCA

DCA (Deep Count Autoencoder) is a deep learning based autoencoder with specialized loss function for scRNA-seq data (Eraslan et al., 2019). In contrast to the traditional autoencoder models that reconstruct the input data itself, DCA defines the reconstruction error as the likelihood of the distribution of the noised model, e.g., zero-inflated negative binomial (ZINB):

$$\begin{aligned} \text{ZINB}(x; \pi, \mu, \theta) &= \pi \delta_0(x) + (1 - \pi) \text{NB}(x; \mu, \theta) \\ \text{Loss} &= \sum_{i=1}^n \sum_{j=1}^p \left( \text{NLL}_{\text{ZINB}}(x_{ij}; \pi_{ij}, \mu_{ij}, \theta_{ij}) + \lambda \pi_{ij}^2 \right) \end{aligned} \quad (2.7)$$

where  $\pi, \mu, \theta$  parameterize the ZINB distribution and  $\text{NLL}_{\text{ZINB}}$  represents the negative log-likelihood of the ZINB distribution. During training, DCA learns feature-specific distribution parameters by minimizing the reconstruction error in an unsupervised manner. The deep learning framework of DCA enables the capturing of the complexity and non-linearity in input data. Furthermore, DCA learns feature-specific parameters mean, dispersion, and dropout probability based on input data. The inferred mean parameter of the distribution represents the denoised reconstruction and can be used for downstream analysis. Figure 2.10d shows the process of the DCA algorithm.



## SCALE

SCALE (Single-Cell ATAC-seq analysis via Latent feature Extraction) is another deep learning-based method specifically developed for scATAC-seq data (Xiong et al., 2019). It combines variational autoencoder (VAE) and Gaussian mixture model (GMM) to model the distribution of high-dimensional scATAC-seq data through the following process:

$$\begin{aligned} p(c) &= \text{Discrete}(c|\pi) \\ p(z|c) &= N(z|\mu_c, \sigma_c^2 \mathbf{I}) \\ p(x|z) &= \text{Bernoulli}(x|\mu_x) \end{aligned} \quad (2.8)$$

where  $c$  is a categorical variable following a discrete distribution,  $p(z|c)$  is a mixture of Gaussian model parameterize by  $\mu_c$  and  $\sigma_c$ , and  $p(x|z)$  is a multi-variable Bernoulli distribution to model the scATAC-seq data. During training, the log-likelihood of the observed scATAC-seq data is maximized:

$$\begin{aligned} \log p(x) &= \log \int_z \sum_c p(x|z) p(z|c) p(c) dz \\ &\geq E_{q(z,c|x)} \left[ \log \frac{p(x,z,c)}{q(z,c|x)} \right] = L_{ELBO}(x) \end{aligned} \quad (2.9)$$

which can be transformed to maximize the evidence lower bound (ELBO). The ELBO can be written with a reconstruction term and a regularization term:

$$L_{ELBO}(x) = E_{q(z,c|x)} [\log p(x|z)] - D_{KL}(q(z,c|x) || p(z,c)). \quad (2.10)$$

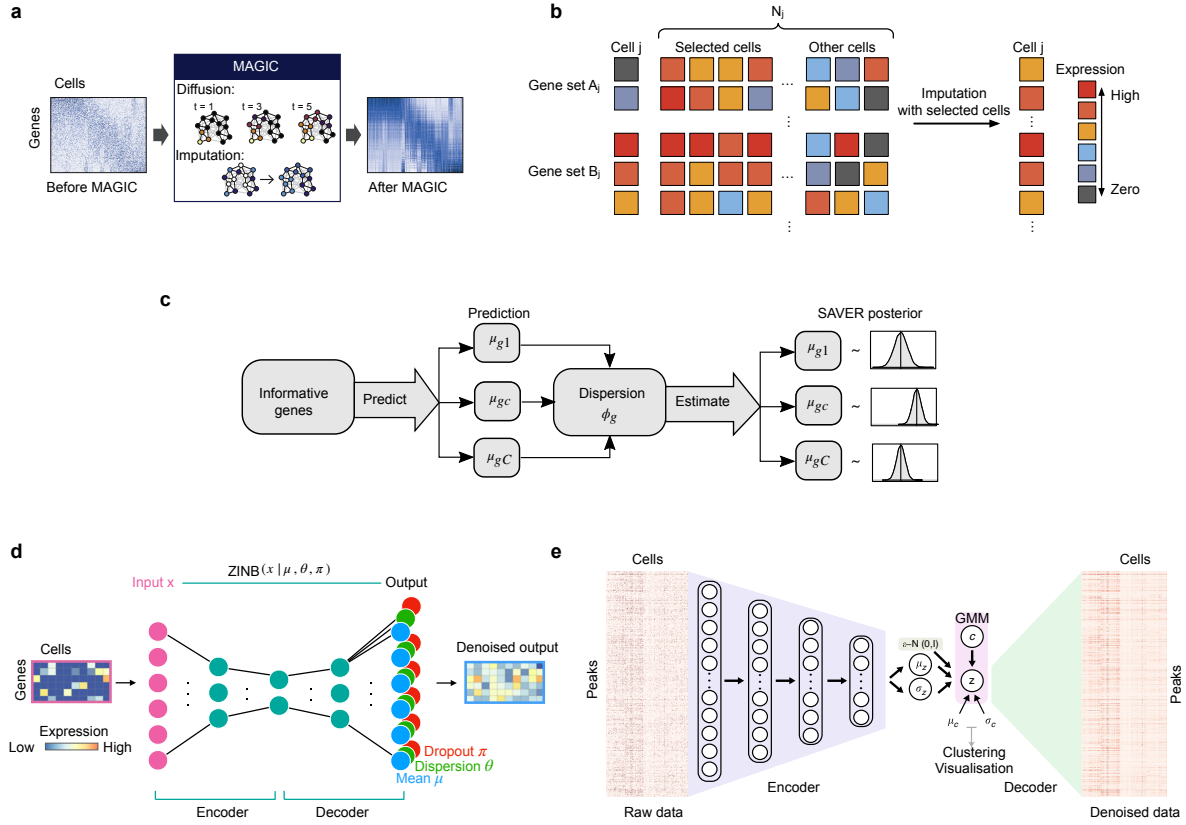
The reconstruction term encourage the imputed data to be similar to the input data, while the regularization term force the latent variable  $z$  to a GMM manifold. Distributions  $q(z,c|x)$  and  $p(x|z)$  are an encoder and a decoder which can be modeled by two neural networks (Figure 2.10e).

## scBFA

scBFA (single-cell Binary Factor Analysis) models the detection pattern observed in the data by ignoring feature quantification measurements and can be applied to both scRNA-seq and scATAC-seq (Li and Quon, 2019). It first creates a binary matrix  $\mathbf{B}$  where each entry represents if a readout is detected for a cell  $i$  ( $i = 1, \dots, N$ ) and a feature  $j$  ( $j = 1, \dots, G$ ). For scRNA-seq input,  $B_{ij} = 1$  indicates that at least one read maps to gene  $j$ , while for scATAC-seq, it means that at least one read maps to locus  $j$  in cell  $i$ . The core idea of scBFA is to explain the high-dimensional detection pattern  $\mathbf{B}$  using two low-dimensional matrices: a  $N \times K$  embedding matrix  $\mathbf{Z}$ , and a  $K \times G$  loading matrix  $\mathbf{A}$ . Subsequently, scBFA defines the following models:

$$\begin{aligned} \text{logit}(\mu_{ij}) &= x_i^T \beta_j + z_i^T \alpha_j + u_i + v_j \\ p(\mathbf{B}; \mathbf{A}, \mathbf{Z}, \beta, \mathbf{X}, \mathbf{u}, \mathbf{v}) &= \prod_{i,j} \text{Bernoulli}(B_{ij} | u_{ij}, \mathbf{A}, \mathbf{Z}, \beta, \mathbf{X}, \mathbf{u}, \mathbf{v}) \end{aligned} \quad (2.11)$$

## 2.4. Related Works



**Figure 2.10: Computational methods for single-cell data denoising and imputation.** **a**, Schematic of MAGIC. The input data consist of a matrix of cells by genes of the data. A cell-to-cell transition matrix is constructed by data diffusion, and imputation is performed using the transition matrix and original data matrix. **b**, Illustration of imputation workflow in scImpute method. scImpute first infers a drop probability for each gene in each cell by fitting a mixture probabilistic model. Next, it imputes the genes with a high drop probability for each cell by borrowing information of the same gene in other similar cells. **c**, Overview of SAVER method. Low-abundance genes are first filtered and only high informative genes are used. Next, SAVER model the observed count of a gene in a cell using a negative binomial random variable through a Poisson-gamma mixture distribution. The goal is to derive the posterior distribution for each gene and cells. **d**, Schematic of DCA with a ZINB loss function. Input is the original count matrix. The blue nodes depict the mean of the negative binomial distribution, which is the main output of the method representing denoised data. The green and red nodes represent the other two parameters of the ZINB distribution, namely dispersion and dropout. **e**, Overview of the SCALE framework. SCALE consists of an encoder and a decoder in the VAE framework. The latent variables are on the GMM manifold parameterized by  $\mu_c$  and  $\sigma_c$ . Source: Van Dijk et al. (2018); Li and Li (2018); Huang et al. (2018); Eraslan et al. (2019); Xiong et al. (2019) (modified to fit thesis format and/or clarify key points). ZINB: zero-inflated negative binomial distribution; VAE: variational auto-encoder; GMM: gaussian mixture model.

where  $\mu_{ij}$  denotes the mean of the Bernoulli distribution determining whether feature  $j$  is detected in cell  $i$  or not,  $\mathbf{A}$  and  $\mathbf{Z}$  represent two low-dimensional matrices used to approximate  $\mathbf{B}$ ,  $u_i$  and  $v_j$

represent the  $i$ th cell-level intercept and  $j$ th feature-specific intercept. Then, scBFA maximizes the following likelihood function:

$$f(\mathbf{B}, \mathbf{A}, \mathbf{Z}, \beta, \mathbf{X}, \mathbf{u}, \mathbf{v}) = \left[ \sum_{i,j} \ln P(B_{ij}; \mathbf{A}, \mathbf{Z}, \beta, \mathbf{X}, \mathbf{u}, \mathbf{v}) \right] - \lambda_1 \|\mathbf{A}\|_2^2 - \lambda_2 \|\mathbf{Z}\|_2^2 - \lambda_3 \|\beta\|_2^2 \quad (2.12)$$

where  $\lambda_1, \lambda_2$ , and  $\lambda_3$  are hyper-parameters for model regularization. After optimization, the two low-dimensional matrices are used for imputation.

Name	Programming	Modality	Task	Technique	Reference
MAGIC	Python/R	scRNA-seq	Denoising	Data smoothing	Van Dijk et al. (2018)
scImpute	R	scRNA-seq	Imputation	Model-based	Li and Li (2018)
SAVER	R	scRNA-seq	Denoising	Model-based	Huang et al. (2018)
DCA	Python	scRNA-seq	Denoising	Deep learning	Eraslan et al. (2019)
scBFA	R	scRNA-seq	Denoising	Matrix factorization	Li and Quon (2019)
SCALE	Python	scATAC-seq	Denoising	Deep learning	Xiong et al. (2019)
cisTopic	R	scATAC-seq	Imputation	Model-based	González-Blas et al. (2019)

**Table 2.1:** Overview of computational denoising and imputation methods.

### 2.4.3 Computational Methods for scATAC-seq Dimensionality Reduction

Dimensionality reduction is another critical problem for scATAC-seq as it forms the foundation for clustering, batch correction, and visualization. Compared to scRNA-seq, generating a meaningful low-dimensional matrix is more challenging due to data sparsity and low count information as previously described (Section 2.4.1). We here first introduce latent semantic indexing (LSI), a widely used method to reduce the dimension of scATAC-seq. Next, we describe cisTopic and SnapATAC, two recent methods designed explicitly for scATAC-seq by considering previously described properties, i.e., high dimensionality, ultra-sparsity, and binarization. Table 2.2 summarizes the aforementioned computational methods.

#### Latent Semantic Indexing

Latent semantic indexing (LSI) is a commonly used approach in text mining field that can simultaneously model the relationships among documents (i.e., observations) based on their constituent words (i.e., variables) and relationships between words based on their occurrence in documents (Deerwester et al., 1990). LSI takes as input a word-document matrix and performs term frequency–inverse document frequency (TF-IDF) transformation. Next, it finds a low-rank approximation to the transformed matrix using singular value decomposition (SVD), in which the  $k$  largest singular values are retained, and the rest set to 0. Afterward, each document and word is represented as a  $k$ -dimensional vector in the space derived by SVD.

This method has been widely used in natural language processing (NLP) for information retrieval since its inception, and Cusanovich et al. (2015) first used it to perform dimensionality reduction for scATAC-seq data by treating cells as documents and peaks as words. Since then, many studies have

## 2.4. Related Works

used it for scATAC-seq data dimensionality reduction (Cusanovich et al., 2018a,b; Satpathy et al., 2019; Domcke et al., 2020). Nevertheless, one challenge of using LSI is determining the optimal number of dimensions for performing SVD, and there is no method to estimate such a number theoretically. Currently, it is a common choice to use 30 dimensions. Another limitation is that the first LSI component is highly correlated to the sequencing depth, indicating it mainly captures technical variation rather than biological variation. In this case, it should be excluded from downstream analysis, thus making LSI less straightforward to use.

### cisTopic

cisTopic is a Bayesian model that can discover *cis*-regulatory topics from scATAC-seq data in an unsupervised manner (González-Blas et al., 2019). It is based on Latent Dirichlet Allocation (LDA), a generative probabilistic model used to discover topics from a collection of documents in NLP (Blei et al., 2003). The input for cisTopic is a region by cell accessibility matrix generated from scATAC-seq data. Then it uses LDA for the modeling of *cis*-regulatory topics. Two distributions are derived from the high-dimensional and sparse scATAC-seq data: i) the probability of a region belonging to a topic (region–topic distribution); ii) the contribution of a topic within a cell (topic-cell distribution). Next, a collapsed Gibbs sampler (Griffiths and Steyvers, 2004) is used to assign each region in each cell to a particular topic by randomly sampling from a distribution where the probability of a region being assigned to a topic is associated with the contribution of that region to the topic and the contribution of that topic to the cell:

$$P(z_i = t | z_{-i}, r) \propto \frac{n_{-i,t}^{(r)} + \beta}{n_{-i,t} + R\beta} \cdot \frac{n_{-i,t}^{(r)} + \alpha}{n_{-i}^c + T\alpha} \quad (2.13)$$

where  $z_i$  is the current assignment to be made,  $z_{-i}$  is the rest of the assignments in the dataset,  $t$  and  $r$  are the given topic and region,  $P(z_i = t | z_{-i}, r)$  is the probability of region  $r$  being assigned to topic  $t$  given the rest of assignments in the dataset,  $n_{-i,t}^{(r)}$  is the number of times that region  $r$  is assigned to topic  $t$  across the dataset without considering the current assignment,  $n_{-i,t}$  is the total number of assignments to topic  $t$  through the dataset,  $\beta$  and  $\alpha$  are Dirichlet hyper-parameters of the prior distribution for the categorical distribution over regions in a topic and over topics in a cell,  $n_{-i}^c$  is the total number of assignments within cell  $c$ ,  $R$  and  $T$  are the number of regions in the dataset and number of topics in the model, respectively. The number of topics can be optimized by selecting the model with the highest log-likelihood. After fitting the model, cisTopic provides two low-dimensional matrices. One is a region-topic matrix containing the contribution of each region to a topic, and another is a topic-cell matrix containing each topic's contribution to a cell. The latter can be used for clustering and visualization as a dimension-reduced matrix for cells. cisTopic is reported to uncover the expected cell types accurately and is more robust compared with LSI, particularly at low sequencing depth (González-Blas et al., 2019). However, this model needs to infer a posterior distribution which has been reported to be computationally expensive (Fang et al., 2021), thus suffering from scalability.

## SnapATAC

SnapATAC is a software package for comprehensively analyzing scATAC-seq data and can process data from up to a million cells (Fang et al., 2021). An essential step in SnapATAC is to reduce the dimension of scATAC-seq data effectively and efficiently. To do so, SnapATAC resolves cellular heterogeneity by directly comparing the similarity in genome-wide accessibility profiles between cells instead of performing matrix decomposition to produce a low-dimension representation as LSI and cisTopic. Specifically, it takes a binary cell-bin matrix  $\mathbf{X} \in \mathbb{R}^{n \times m}$  as input where  $n$  represents the number of cells and  $m$  represents bins. First, it constructs a  $\mathbf{J} \in \mathbb{R}^{n \times n}$  similarity matrix. The similarity between cells is calculated using the Jaccard coefficient, which is defined as the size of the intersection divided by the size of the union:

$$Jaccard(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|}. \quad (2.14)$$

Because the Jaccard similarity between cells is highly correlated to the reads depth (Fang et al., 2021), SnapATAC-seq next fits a regression model and normalizes the observed Jaccard coefficient matrix. Using this normalized similarity matrix  $\mathbf{N}$ , SnapATAC then produces a low-dimension matrix using a diffusion map (DM) (Coifman et al., 2005) by:

$$\begin{aligned} \mathbf{A} &= \mathbf{D}^{-1/2} \mathbf{N} \mathbf{D}^{-1/2} \\ \mathbf{A} &= \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \end{aligned} \quad (2.15)$$

where  $\mathbf{D} \in \mathbb{R}^{n \times n}$  is a diagonal matrix composed as  $D_{i,i} = \sum_j N_{i,j}$ , and  $\mathbf{U} \in \mathbb{R}^{n \times n}$  are a set of eigenvectors, the diagonal matrix  $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$  has the eigenvalues in descending order as its entries. Finally, the first  $k$  eigenvectors are used as the low-dimensional representation.

Name	Programming	Task	Technique	Reference
LSI	R	Dim. Reduction	SVD	Cusanovich et al. (2015)
cisTopic	R	Dim. Reduction/Imputation	LDA	González-Blas et al. (2019)
SnapATAC	R	Dim. Reduction	DM	Fang et al. (2021)

**Table 2.2:** Overview of computational methods for scATAC-seq dimensionality reduction.

## 2.5 Discussion

In this chapter, we first introduced the basic concepts of DNA, chromatin organization, and accessibility from a biological perspective. Then, we described ATAC-seq to measure chromatin accessibility at bulk and single-cell resolution. Next, we presented the standard computational workflow for analyzing bulk and single-cell ATAC-seq data. Finally, we discussed the computational challenges and reviewed the recent development for analyzing scATAC-seq data in the field. In particular, we focused on two computational problems: data imputation and dimensionality reduction.

## 2.5. Discussion

In scATAC-seq, Tn5 insertion generates a maximum of 2 fragments per cell in a small (200 bp) chromatin-accessible region. Subsequent steps of the ATAC-seq protocol cause a loss of a large proportion of these fragments. For example, only DNA fragments with two different Tn5 adapters, which are statistically present in 50% of the fragment, are amplified in the PCR step (Buenrostro et al., 2015a). Further material losses occur during single-cell isolation, liquid handling, reads capture, or by simple financial restrictions of sequencing depth.

Usually, the first step for analysis of scATAC-seq is the detection of OC regions by calling peaks on the scATAC-seq library by ignoring cell information. Next, a matrix is built by counting the number of digestion events per cell in each of the previously detected regions. This matrix usually has a very high dimension (up to  $10^6$  regions) and a maximum of two digestion events are expected for a region per cell. These characteristics render the scATAC-seq count matrix sparse, i.e. 3% of non-zero entries. In contrast, scRNA-seq have less severe sparsity ( $>10\%$  of non-zeros) than scATAC-seq due to smaller dimension ( $< 20,000$  genes for mammalian genomes) and lower dropout rates for genes with high or moderate expression levels, as shown in Figure 2.9. This sparsity poses challenges in the identification of cell-specific OC regions and is likely to affect downstream analysis as clustering and detection of regulatory features.

Although several computational methods have been developed to address this issue for scRNA-seq data (e.g., MAGIC, scImpute, SAVER, DCA, and scBFA, as described in Section 2.4.2), these methods were not designed to deal with the sparse of scATAC-seq data. Moreover, they assumed that the data followed a specific distribution, e.g., scImpute used Gamma-Normal distribution (Equation 2.1), SAVER used Gamma-Poisson mixture (Equation 2.3), and DCA used zero-inflated negative binomial distribution (Equation 2.7), making them inappropriate for scATAC-seq due to the low count nature of scATAC-seq data. Until date, there are only two approaches for imputation methods for scATAC-seq data (SCALE and cisTopic). However, SCALE, a deep learning based method, requires a graphics processing unit (GPU) for training. The usual small size of GPU memory limits the number of cells to be analyzed. cisTopic is a Bayesian-based method, which was reported to have an exponential increase of the running time for an increasing number of reads (Chen et al., 2019). Therefore, both approaches are likely to have scalability issues with large data sets.

Dimensionality reduction represents another unmet need for scATAC-seq data analysis. Currently, there are only three methods available for this task (i.e., LSI, cisTopic, and SnapATAC), compared with the vast number of tools for scRNA-seq (Sun et al., 2019). One reason that makes this task difficult is because the matrix generated by scATAC-seq is binary where only two possible values are presented, i.e., 1 means accessible and 0 means non-accessible. The commonly used approaches for scRNA-seq dimensionality reduction, such as PCA, cannot directly deal with this binary matrix. Therefore, more specific methods are needed in this regard.

In addition to methods development, the systematic evaluation of imputation and dimensionality reduction for scATAC-seq data is still an open problem. Although a recent study has benchmarked several computational methods for scATAC-seq on a number of synthetic and real datasets from different assays (Chen et al., 2019), the results are solely based on clustering of cells which might be biased by using different algorithms. Furthermore, the clustering is an indirect metric for matrix imputation, and it does not reflect the true performance of an imputation method, i.e., what fraction of

the missing open chromatin regions are recovered after imputation. Moreover, it is still unclear if the downstream analysis of scATAC-seq can also be benefited by imputation. Altogether, an unbiased and systematic evaluation with different metrics for scATAC-seq data imputation and dimensionality reduction methods is required.

In this thesis, we investigate methods for imputing and reducing dimensions of scATAC-seq data. Our goals are summarized as follows:

- The development of a new computational method that is able to accurately quantify single-cell chromatin accessibility status by data denoising and imputation. Moreover, this method should also provide a low-dimensional representation of the raw data. This dimension reduced matrix will be used for clustering, batch correction, and data visualization when dealing with large-scale scATAC-seq. Furthermore, we will generate simulated scATAC-seq data and explore how to perform model selection.
- The evaluation of available imputation methods under the context of scATAC-seq. We will apply the previously described methods (Section 2.4.2) to impute several real-world scATAC-seq data sets with available true labels. The performance of each method will be evaluated based on cell clustering, peaks recovery and cell-to-cell similarity estimation. Moreover, we will also benchmark the scalability of each method, i.e., the required CPU memory and running time.
- The evaluation of available dimensionality reduction methods. We will generate the low-dimensional matrix using our approach and the previously described algorithms (Section 2.4.3). The results will be evaluated based on clustering and distance accuracy.
- The investigation of the impact of imputation on downstream analysis of scATAC-seq data. We will test whether the scATAC-seq computational pipelines will benefit using the imputed matrix, as the sparsity has been eliminated.
- The application of our method to novel scATAC-seq data. In collaboration with Rafael Kra-  
mann and Christoph Kuppe in the Institute of Experimental Medicine and Systems Biology, RWTH Aachen University Medical School, we plan to perform scATAC-seq on whole mouse kidney tissues in homeostasis and at two time points (day two and day ten) after injuring with fibrosis. We will also evaluate our approach in its power to detect cells in such a complex disease dataset by using an independent scRNA-seq dataset generated from the same model as gold standard. We will also seek to obtain new biological insights about this process.





## Methods

In the previous chapter, we introduced the concepts of ATAC-seq to profile chromatin accessibility at bulk and single-cell resolution. We also described the standard computational workflow for the analysis of bulk and single-cell ATAC-seq data. One of the main challenges to analyze the scATAC-seq data is the sparsity and high-dimensionality. Therefore, this thesis aims to develop a new computational method for scATAC-seq data imputation and dimensionality reduction.

In this chapter, we exclusively present and formalize our computational solution towards this goal. Specifically, we first introduce the notation that is necessary to formalize our method (Section 3.1). Next, we describe a two-step strategy to normalize the input data, i.e., data binarization and transformation (Section 3.2). Then, we present our approach for data imputation and dimensionality reduction. Moreover, we introduce a computational strategy to automatically determine the number of components in the model (Section 3.3). Next, we describe the implementation details of the methods (Section 3.4). Finally, we close this chapter with a few concluding remarks on the methodology choice and novelty of our approach (Section 3.5).

### 3.1 Notation

Throughout this chapter, we denote matrices by boldface capital letters (e.g.,  $\mathbf{A}$ ), denote vectors by boldface lowercase letters (e.g.,  $\mathbf{a}$ ), denote scalars by lowercase letters (e.g.,  $a$ ), denote the  $i$ th entry of a vector  $\mathbf{a}$  by  $a_i$ , denote the  $i$ th row of a matrix  $\mathbf{A}$  by  $\mathbf{a}_i$ , denote the  $i$ th column of a matrix  $\mathbf{A}$  by  $\mathbf{a}_{:i}$ , denote an element of  $(i, j)$  of a matrix  $\mathbf{A}$  by  $a_{ij}$ , and denote non-negative real numbers by  $\mathbb{R}_+$ . More specifically, we define the following notation:

- $\mathbf{X} \in \mathbb{R}_+^{m \times n}$ , the peak-by-cell count matrix;
- $m$ , the total number of peaks (i.e., variables);
- $n$ , the total number of cells (i.e., observations);
- $i \in \{1, \dots, m\}$ , the index of peaks;
- $j \in \{1, \dots, n\}$ , the index of cells;
- $\mathbf{B} \in \mathbb{R}_+^{m \times n}$ , the binarized matrix of  $\mathbf{X}$ ;
- $\mathbf{T} \in \mathbb{R}_+^{m \times n}$ , the term frequency matrix;
- $\mathbf{d} \in \mathbb{R}_+^m$ , the vector of inverse document frequency;
- $\mathbf{Y} \in \mathbb{R}_+^{m \times n}$ , the TF-IDF transformed matrix;

### 3.2. Data Normalization

- $\mathbf{M} \in \mathbb{R}_+^{m \times n}$ , the L2 normalized matrix of  $\mathbf{Y}$ ;
- $\mathbf{W} \in \mathbb{R}_+^{m \times k}$ , the low-dimensional representation of peaks;
- $\mathbf{H} \in \mathbb{R}_+^{k \times n}$ , the low-dimensional representation of cells;
- $1 \leq k \leq \min(m, n)$ , the number of components.

## 3.2 Data Normalization

---

We first perform data normalization for the input count matrix which is generated as described in Section 2.3.2. This procedure includes two steps: (i) data binarization (Section 3.2.1); (ii) data transformation (Section 3.2.2). We describe each step in detail below.

### 3.2.1 Binarization

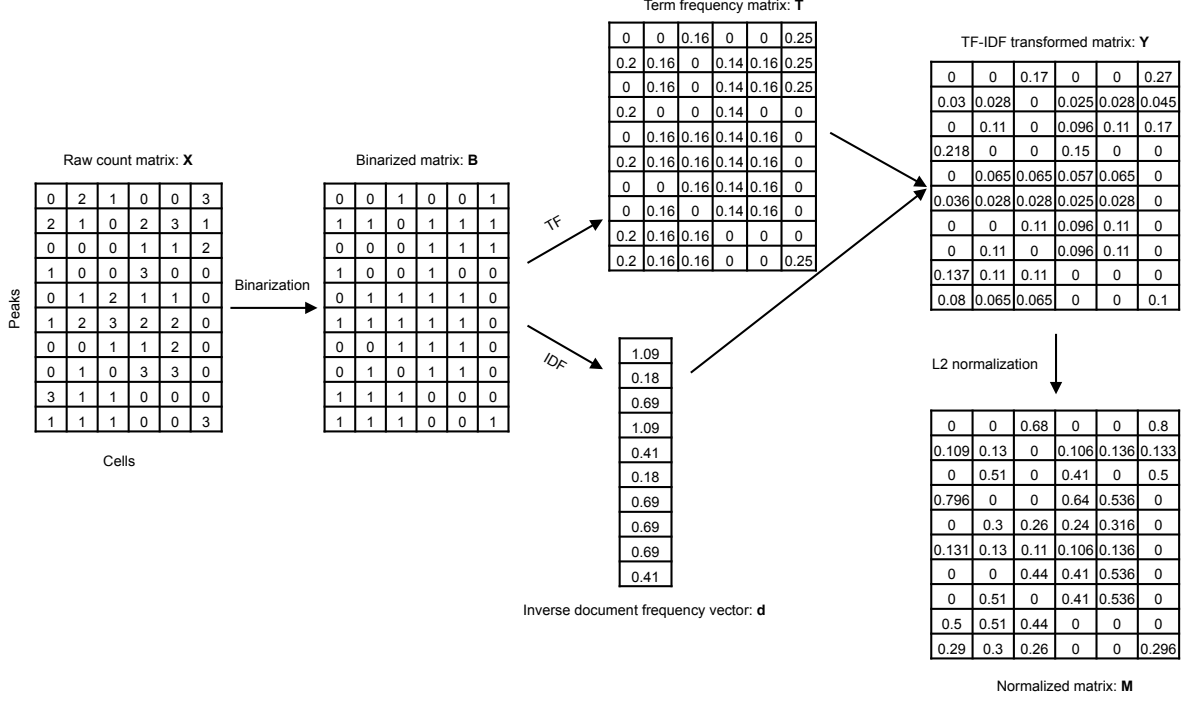
As mentioned in Section 2.4.1, for scATAC-seq data, a particular chromatin site in a single cell can either be accessible or non-accessible, and a maximum of two Tn5 digestion events are expected due to limited DNA copies. Consequently, the differences in reads count may largely reflect technical effect (e.g., reads amplification bias) rather than true biological variation. Therefore, we here create a binary accessibility matrix  $\mathbf{B}$  to eliminate the potential technical bias by the following operation:

$$b_{ij} = \begin{cases} 1 & x_{ij} > 0 \\ 0 & x_{ij} = 0 \end{cases} \quad (3.1)$$

where 1 indicates the peak  $i$  is accessible in cell  $j$ , and 0 could indicate non-accessible or non-measured (i.e., dropout or missing values). It is worth pointing out that these two inferences represent very different biological standpoints. Because of this, the 1s contain information, and the 0s do not (Granja et al., 2021).

### 3.2.2 Transformation

We next perform data transformation to remove further technical factors, including the number of detected peaks in each cell, which are usually caused by different reaction efficiency and can mix up a technical variation with biological heterogeneity. To do so, we apply a statistical technology, called term frequency-inverse document frequency (TF-IDF), to transform the binary accessibility matrix  $\mathbf{B}$  to a TF-IDF transformed matrix  $\mathbf{Y}$ . TF-IDF technique was originally developed in the nature language processing (NLP) field to access document similarity based on counts of the word (Salton and McGill, 1986). This method has been widely used for information retrieval and text mining. The intuition of applying such an approach to scATAC-seq data transformation is that, in the context of scATAC-seq, we can consider the cells as the *documents* and the accessible regions/peaks as the *words*. The transformed value in matrix  $\mathbf{Y}$  generally reflects how important a peak is to a specific cell.



**Figure 3.1: A toy example of scATAC-seq data normalization.** This figure demonstrates the data normalization process using a toy accessibility count matrix. First, to remove the potential reads amplification bias, the fragment count matrix **X** is binarized to generate a binary matrix **B**. Next, to further remove technical effects (e.g., sequencing depth), a term frequency matrix **T** and an inverse document frequency vector **d** are calculated, respectively. A TF-IDF transformed matrix **Y** is obtained by multiplying the vector **d** with each column in matrix **T**. Finally, the L2 normalization approach is used to further normalize the data for each cell, and the normalized matrix **M** is used for imputation and dimensionality reduction.

To perform data transformation, we first calculate a term frequency (TF) matrix **T** by:

$$t_{i,j} = \frac{b_{ij}}{\sum_{m'=1}^m b_{m'j}} \quad (3.2)$$

where  $t_{i,j}$  ranges from 0 and 1, representing the importance of a peak  $i$  in cell  $j$ , and  $b_{ij}$  is obtained from the binary accessibility matrix **B** by using Equation 3.1. As a way of normalization, this process aims to remove the sequencing depth bias (i.e., number of detected peaks bias) which is usually caused by stochastic molecular sampling during sequencing. More specifically, if a cell has more peaks detected than other cells, the weight of each peak in this cell will be proportionally decreased.

Next, we compute an inverse document frequency (IDF) vector **d** by:

$$d_i = \ln\left(\frac{n}{\sum_{j=1}^n b_{ij}}\right) \quad (3.3)$$

where  $n$  indicates the total number of cells and  $\ln$  represent the natural logarithm. This term is a non-negative value, typically representing the importance of a peak across all cells. Intuitively, if

### 3.3. Data Imputation and Dimensionality Reduction

a peak is common (i.e., it is detected in most cells), it might be less important because it does not provide too much information to distinguish relevant and non-relevant cells. Therefore, the weight of this peak should be scaled down. On the other hand, if a peak is rare, it could be a critical feature for identifying rare cell populations, so the weight of this peak should be scaled up.

Subsequently, we calculate the TF-IDF transformed matrix  $\mathbf{Y}$  as the product of corresponding TF and IDF score:

$$y_{i,j} = t_{i,j} \cdot d_i. \quad (3.4)$$

Finally, we use the L2 normalization technique to further normalize the data for each cell as follows:

$$m_{ij} = \frac{y_{i,j}}{\sqrt{\sum_{m'=1}^m (y_{m',j})^2}}. \quad (3.5)$$

This matrix  $\mathbf{M}$  represents a normalized accessibility matrix which is later used for imputation and dimensionality reduction. Figure 3.1 shows a toy example of scATAC-seq data normalization process.

## 3.3 Data Imputation and Dimensionality Reduction

---

Having obtained the normalized matrix  $\mathbf{M}$ , we next would like to learn a low-dimensional representation for the cells by reducing the dimensionality of the data through a mapping function. Moreover, an imputed matrix recovering the missing signal is also expected. Given that all the elements of  $\mathbf{M}$  are non-negative, we here choose to use the NMF technique for this task. In this section, we first give a brief introduction to NMF and formalize our problem (Section 3.3.1). Next, we describe a numerical method to solve the optimization problem (Section 3.3.2). Finally, we introduce an approach to automatically determine the number of components, a hyper-parameter in the NMF model (Section 3.3.4).

### 3.3.1 NMF

Non-negative matrix factorization (NMF) was first introduced in 1994 as a new variant of factor analysis (Paatero and Tapper, 1994). The problem can be formalized as given a non-negative matrix  $\mathbf{V}$ , find matrix factors  $\mathbf{W}$  and  $\mathbf{H}$  such that:

$$\mathbf{V} \approx \mathbf{WH}, \quad s.t. \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0 \quad (3.6)$$

where  $\mathbf{W}$  and  $\mathbf{H}$  are non-negative and usually have a lower rank than the original matrix  $\mathbf{V}$ . Solving this problem results in a compressed version of the original data matrix. In contrast to other dimensionality reduction methods (e.g., PCA), NMF was showed to be able to learn the parts of objects and provide easily interpretable features because of its non-negativity constraints (Lee and Seung, 1999). Furthermore, the authors provided a simple and efficient algorithm to solve the optimization problem (Lee and Seung, 2001). Since then, the NMF technique has been widely applied to a number of real-world applications for the analysis of high-dimensional data, such as image process (Guillamet et al., 2003), text mining (Dhillon and Sra, 2005), audio signal processing (Gemmeke et al., 2013),

and bioinformatics (Welch et al., 2019; Shiga et al., 2020).

To apply NMF to factorize the normalized matrix  $\mathbf{M}$  into two smaller matrices, we define the following model:

$$\mathbf{M} = \mathbf{WH} + \mathbf{Z} \quad (3.7)$$

where  $\mathbf{W}$  is a  $m$  by  $k$  matrix containing the factor values for peaks,  $\mathbf{H}$  is a  $k$  by  $n$  matrix containing the factor values for cells.  $\mathbf{Z} \in \mathbb{R}^{m \times n}$  represents some additive noise present in the matrix  $\mathbf{M}$ , and we model it as independently and identically (i.i.d) Gaussian noise:

$$z_{ij} \sim N(0, \sigma_{ij}^2) \quad (3.8)$$

where  $\sigma_{ij}$  represents the variance. Figure 3.2 shows the conceptual illustration of the model. More specifically, each element in  $\mathbf{M}$  can be written as:

$$m_{ij} = \sum_{k'=1}^k w_{ik'} \cdot h_{k'j} + z_{ij}. \quad (3.9)$$

Our goal is to find out  $\mathbf{W}$  and  $\mathbf{H}$  that can construct  $\mathbf{M}$  with the lowest error. For this, we use the sum of square errors to measure construction quality and define the following objective function with respect to  $\mathbf{W}$  and  $\mathbf{H}$ :

$$f(\mathbf{W}, \mathbf{H}) = \underbrace{\frac{1}{2} \|\mathbf{M} - \mathbf{WH}\|^2}_{\text{error}} + \underbrace{\frac{\lambda}{2} \|\mathbf{W}\|^2 + \frac{\lambda}{2} \|\mathbf{H}\|^2}_{\text{regularization}}, \quad s.t. \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0 \quad (3.10)$$

where  $\|\cdot\|^2$  is the Frobenius norm of a matrix define as:

$$\|\mathbf{W}\|^2 = \sum_{i=1}^m \sum_{k'=1}^k w_{ik'}^2. \quad (3.11)$$

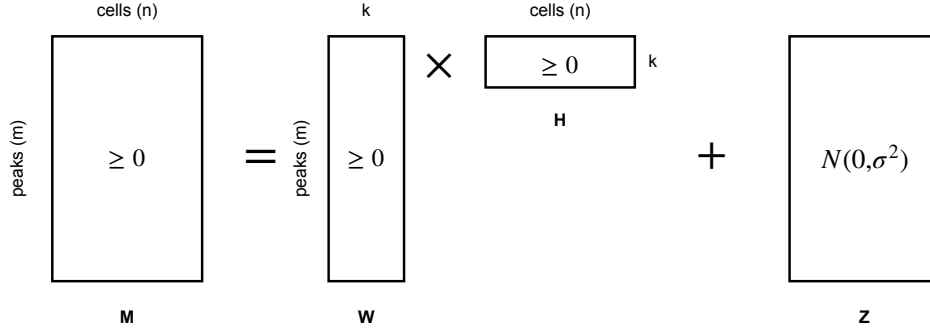
The first item is the estimator of square loss for each element in  $\mathbf{M}$ , and the second item is the regularization which prevent the model from over-fitting. The parameter  $\lambda$  controls the balance between reconstruction error and model regularization. We will evaluate the impact of this parameter on the final results in (Section 5.1.1). Finally,  $\mathbf{W}$  and  $\mathbf{H}$  are found by minimizing the above objective function:

$$\mathbf{W}, \mathbf{H} = \arg \min_{\mathbf{W}, \mathbf{H}} f(\mathbf{W}, \mathbf{H}) \quad s.t. \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0. \quad (3.12)$$

### 3.3.2 Solving the Optimization Problem

It is worth pointing out that this optimization problem was proved to be non-convex, ill-posed, and NP-hard (Vavasis, 2010). This means that it is unlikely to find an optimal global resolution in a reasonable computation time. Fortunately, several heuristic algorithms have been proposed to find out the local optimal, including multiplicative-update algorithm (Lee and Seung, 2001), projected gradient methods (Lin, 2007), active set method (Kim and Park, 2008), and coordinate descent method (Ci-

### 3.3. Data Imputation and Dimensionality Reduction



**Figure 3.2: Conceptual illustration of the defined NMF model.** The normalized matrix  $\mathbf{M}$  is factorized into two matrices  $\mathbf{W}$ ,  $\mathbf{H}$ , and a noise matrix  $\mathbf{Z}$  modeled as Gaussian noise. Both  $\mathbf{W}$  and  $\mathbf{H}$  have lower rank than  $\mathbf{M}$  and are non-negative. Therefore, these two low-rank matrices are considered as a compression of the original matrix  $\mathbf{M}$ .

chocki and Phan, 2009). Among them, coordinate descent is regarded as one of the state-of-the-art techniques for solving the problem. The idea is that the objective function can be minimized along one direction at one time. Specifically, at each iteration, one of the two factors is fixed, and the other is updated in such a way that the objective function is reduced:

$$(\mathbf{W}^0, \mathbf{H}^0) \rightarrow (\mathbf{W}^1, \mathbf{H}^0) \rightarrow (\mathbf{W}^1, \mathbf{H}^1) \rightarrow \dots \rightarrow (\mathbf{W}^{T_{max}}, \mathbf{H}^{T_{max}}) \quad (3.13)$$

where  $\mathbf{W}^0$  and  $\mathbf{H}^0$  represent the initial matrices, and  $T_{max}$  represents the maximum number of iterations. This amounts to a two-block coordinate descent method.

Moreover, for each factor, the problem can be further simplified as a simple univariate quadratic problem if we optimize one single variable at a time with all others fixed. For example, consider minimizing the function over  $w_{it}$  with all other elements in  $\mathbf{W}$  fixed, then the objective function with respect to  $w_{it}$  can be written as follows:

$$\begin{aligned} f(w_{it}) &= \frac{1}{2} \sum_{j=1}^n (m_{ij} - \sum_{k' \neq t}^k w_{ik'} h_{k'j})^2 + \frac{\lambda}{2} w_{it}^2 \\ &= \frac{1}{2} \sum_{j=1}^n (m_{ij} - \sum_{k' \neq t}^k w_{ik'} h_{k'j} - w_{it} h_{tj})^2 + \frac{\lambda}{2} w_{it}^2 \quad s.t. \quad w_{it} \geq 0. \end{aligned} \quad (3.14)$$

The gradient of  $f$  with respect to  $w_{ij}$  is written as:

$$\begin{aligned} \nabla f(w_{ij}) &= \sum_{j=1}^n (m_{ij} - \sum_{k' \neq t}^k w_{ik'} h_{k'j} - w_{it} h_{tj}) (-h_{tj}) + \lambda w_{it} \\ &= \sum_{j=1}^n (w_{it} h_{tj}^2 + h_{tj} \sum_{k' \neq t}^k w_{ik'} h_{k'j} - m_{ij} h_{tj}) + \lambda w_{it} \\ &= (\sum_{j=1}^n w_{it} h_{tj} + \lambda w_{it}) + \sum_{j=1}^n (h_{tj} \sum_{k' \neq t}^k w_{ik'} h_{k'j} - m_{ij} h_{tj}) \\ &= w_{it} (\sum_{j=1}^n h_{tj}^2 + \lambda) - \sum_{j=1}^n (m_{ij} h_{tj} - h_{tj} \sum_{k' \neq t}^k w_{ik'} h_{k'j}). \end{aligned} \quad (3.15)$$

To minimize the function, we set the gradient to zero:

$$\nabla f(w_{ij}) = w_{it} \left( \sum_{j=1}^n h_{tj}^2 + \lambda \right) - \sum_{j=1}^n (m_{ij} h_{tj} - h_{tj} \sum_{k' \neq t}^k w_{ik'} h_{k'j}) = 0. \quad (3.16)$$

Then the optimal solution is found as:

$$w_{it} = \frac{\sum_{j=1}^n (m_{ij} h_{tj} - h_{tj} \sum_{k' \neq t}^k w_{ik'} h_{k'j})}{\sum_{j=1}^n h_{tj}^2 + \lambda}. \quad (3.17)$$

Given the non-negativity constraint of  $w_{it}$ , the final solution is written as follows:

$$w_{it} = \max\left(0, \frac{\sum_{j=1}^n (m_{ij} h_{tj} - h_{tj} \sum_{k' \neq t}^k w_{ik'} h_{k'j})}{\sum_{j=1}^n h_{tj}^2 + \lambda}\right). \quad (3.18)$$

We iteratively apply the above update rule for all elements in  $\mathbf{W}$ . Of note, the original objective function remains unchanged when the matrices  $\mathbf{W}$  and  $\mathbf{H}$  are transposed, i.e.,

$$\begin{aligned} f(\mathbf{W}, \mathbf{H}) &= \frac{1}{2} \|\mathbf{M} - \mathbf{WH}\|^2 + \frac{\lambda}{2} \|\mathbf{W}\|^2 + \frac{\lambda}{2} \|\mathbf{H}\|^2 \\ &= \frac{1}{2} \|\mathbf{M}^T - \mathbf{H}^T \mathbf{W}^T\|^2 + \frac{\lambda}{2} \|\mathbf{W}^T\|^2 + \frac{\lambda}{2} \|\mathbf{H}^T\|^2 \\ &= f(\mathbf{H}^T, \mathbf{W}^T). \end{aligned} \quad (3.19)$$

Based on this observation, we can update  $\mathbf{H}$  using the same rule:

$$h_{tj} = \max\left(0, \frac{\sum_{i=1}^m (m_{ij} w_{it} - w_{it} \sum_{k' \neq t}^k h_{jk'} w_{k'i})}{\sum_{i=1}^m w_{it}^2 + \lambda}\right). \quad (3.20)$$

In order to make the results more robust, in each step of optimizing the factor, we randomly shuffle the order of cells in the matrix. The above iteration is carried out until a termination criterion is met, e.g., number of iteration performed. Afterward, we multiply the matrix  $\mathbf{W}$  and  $\mathbf{H}$  to obtain the imputed matrix:

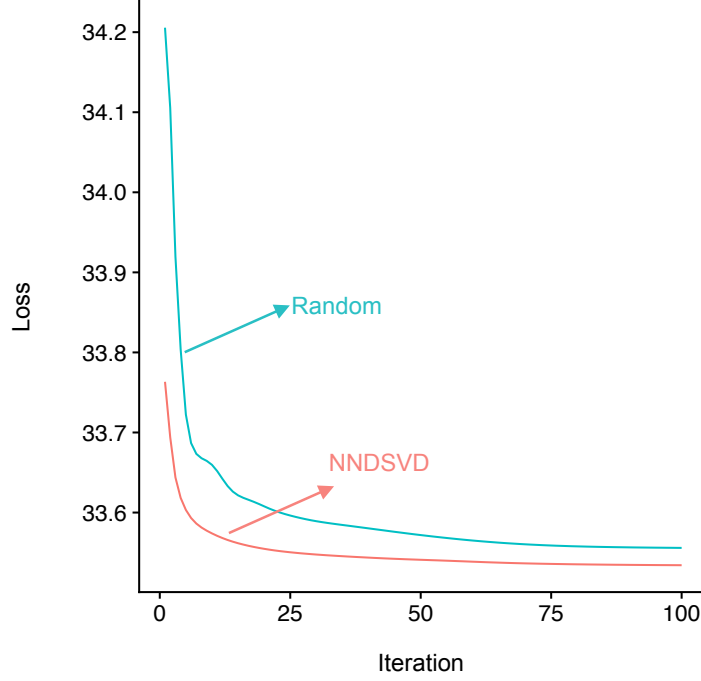
$$\hat{\mathbf{M}} = \mathbf{WH} \quad (3.21)$$

and use  $\mathbf{H}$  as a dimension reduced matrix for cells.

### 3.3.3 Initialization

The optimization procedure starts with some initial values of matrices  $\mathbf{W}$  and  $\mathbf{H}$  with only non-negative elements. The simplest way of obtaining such initial values is to generate two random non-negative matrices. However, this approach suffers from slow convergence and requires more iterations to achieve an optimal solution. We, therefore, use a more effective method, called non-negative double singular value decomposition (NDSVD), to generate the initial values for  $\mathbf{W}$  and  $\mathbf{H}$  based on two processes of SVD (Boutsidis and Gallopoulos, 2008).

### 3.3. Data Imputation and Dimensionality Reduction



**Figure 3.3: Convergence curve of the model with different initialization approaches** This figure compares the convergence rate of the optimization procedure using randomly initialized values or NNDSVD-based initialization. The x-axis represents the number of iterations, and the y-axis represents the estimated loss. The results are based on a scATAC-seq matrix with 125,647 peaks and 1,224 cells.

More formally, the input matrix  $\mathbf{M}$  is first factorized by SVD as:

$$\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^T \quad (3.22)$$

where  $\mathbf{U} \in \mathbb{R}^{m \times m}$  and  $\mathbf{V} \in \mathbb{R}^{n \times n}$  are two unitary square matrices with  $\mathbf{u}_{:i}$  and  $\mathbf{v}_{:i}$  representing the left and right singular vectors,  $\mathbf{S} \in \mathbb{R}_+^{m \times n}$  is a rectangular diagonal matrix with  $s_{ii}$  representing the singular values of  $\mathbf{M}$ . Without loss of generality, we assume that  $s_{11} \geq s_{22} \geq \dots \geq s_{rr} > 0$ , where  $r$  represents the index of the lowest singular value. Then, for a rank  $k \leq r$ ,  $\mathbf{M}$  can be optimally approximated by:

$$\mathbf{M} \approx \sum_{k'=1}^k \mathbf{C}_{i'} \approx \sum_{k'=1}^k \sum_k s_{k'k'} \mathbf{u}_{:k'} \mathbf{v}_{:k'}^T. \quad (3.23)$$

Suppose we want to generate two matrices  $\mathbf{W} \in \mathbb{R}_+^{m \times k}$  and  $\mathbf{H} \in \mathbb{R}_+^{k \times n}$ , according to the Perron-Frobenius theorem, because  $\mathbf{M}$  is non-negative, its maximum left and right singular vectors are also guaranteed to be non-negative. Therefore, the first column of  $\mathbf{W}$  and the first row of  $\mathbf{H}$  can be initialized as:

$$\begin{aligned} \mathbf{m}_{:1} &= \sqrt{s_{11}} \cdot \mathbf{u}_{:1} \\ \mathbf{h}_{1:} &= \sqrt{s_{11}} \cdot \mathbf{v}_{:1}^T. \end{aligned} \quad (3.24)$$

For other columns and rows of matrices  $\mathbf{W}$  and  $\mathbf{H}$ , a similar approach is used. Specifically, for any



index  $2 \leq k' \leq k$ , a matrix  $C_{k'}$  is obtained by:

$$C_{k'} = s_{k'k'} u_{:k'} v_{:k'}^T. \quad (3.25)$$

We can zero out all negative elements of  $C_{k'}$  to obtain a non-negative matrix  $C_{k'}^+$  and use the above process to initialize the  $k'$  column and row of  $\mathbf{W}$  and  $\mathbf{H}$ . Figure 3.3 shows the convergence curves by using either randomly initialized matrices or NNDSVD-based initialization for model optimization. Finally, by combining the above described steps, we here propose a new computation method for scATAC-seq data imputation and dimensionality reduction. The algorithm is formalized in Algorithm 3.1.

---

**Algorithm 3.1:** Imputation and dimensionality reduction for scATAC-seq data

---

**Input** :  $\mathbf{X}$ , a  $m \times n$  matrix containing the number of reads for all peaks and cells;  
 $m$ , the number of peaks;  
 $n$ , the number of cells;  
**Parameter:**  $k$ , the number of components;  
 $\lambda$ , the regularization parameter;  
 $T_{max}$ , the maximum of iteration;  
**Output** :  $\mathbf{W}$ , a  $m \times k$  matrix of low-dimensional representation of peaks;  
 $\mathbf{H}$ , a  $k \times n$  matrix of low-dimensional representation of cells;  
 $\hat{\mathbf{M}}$ , a  $m \times n$  matrix containing imputed data for all peaks and cells;

```

1  $\mathbf{B} = \text{BINARIZATION}(\mathbf{X})$  ; // Section 3.2.1
2  $\mathbf{M} = \text{TRANSFORMATION}(\mathbf{B})$  ; // Section 3.2.2
3  $\mathbf{W}, \mathbf{H} = \text{INITIALIZATION}(\mathbf{M})$  ; // Section 3.3.3
4 while  $iter < T_{max}$  do
5   for  $t \leftarrow 1$  to  $k$  by 1 do
6     for  $i \leftarrow 1$  to  $m$  by 1 do
7       | Update  $w_{it}$  using Equation 3.18 ; // Update  $\mathbf{W}$ 
8     end
9   end
10  for  $t \leftarrow 1$  to  $k$  by 1 do
11    for  $j \leftarrow 1$  to  $n$  by 1 do
12      | Update  $h_{tj}$  using Equation 3.20 ; // Update  $\mathbf{H}$ 
13    end
14  end
15 end
16  $\hat{\mathbf{M}} = \mathbf{W} \mathbf{H}$ ;
17 return  $\mathbf{W}, \mathbf{H}, \hat{\mathbf{M}}$ 

```

---

### 3.3.4 Determining the Hyper-parameters

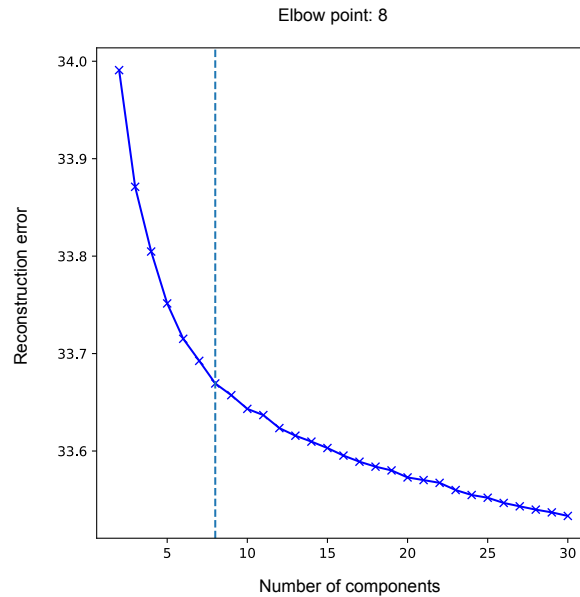
There are two hyper-parameters in our proposed method, i.e., regularization parameter  $\lambda$  and number of components  $k$ . The parameter  $\lambda$  is used to prevent the model from overfitting, and it is empirically evaluated on a simulated scATAC-seq dataset to test its impact on various downstream tasks (Section 4.1.2). The number of components  $k$  determines the intrinsic dimensions of a matrix and thus

### 3.4. Implementation

is highly dataset-specific. Although a higher  $k$  is able to better approximate the original matrix, it is computationally expensive. On the other hand, a lower  $k$  may lead to underfitting. In order to automatically select an appropriate  $k$  for a given dataset, we input a number of different values for  $k$ , e.g., from 2 to 30. For each  $k$ , we run Algorithm 3.1 to obtain the imputed matrix  $\hat{\mathbf{M}}$  and compute the reconstruction error as:

$$e_k = ||M_k - \hat{\mathbf{M}}_k||^2. \quad (3.26)$$

Next, we build an error curve and detect the elbow point of this curve using the Kneedel algorithm (Satopaa et al., 2011). The elbow point is roughly defined as the point of maximum curvature in a system. It represents the best balance between the cost and the expected performance benefit, in our case., the number of components  $k$  and the reconstruction error  $e$ . Figure 3.4 shows an example of using this approach to determine the number of components for a given scATAC-seq dataset composed of 1,224 cells from 8 different cell types. The algorithm correctly detects  $k = 8$  as the elbow point, corresponding to the number of cell types in the data.



**Figure 3.4: Estimation of the number of components using elbow detection algorithm** This figure shows an example of using elbow detection algorithm to estimate the optimal number of components. The x-axis represents the number of components, and the y-axis represents the reconstruction error estimated using Equation 3.26. The dashed line indicates the detected elbow point.

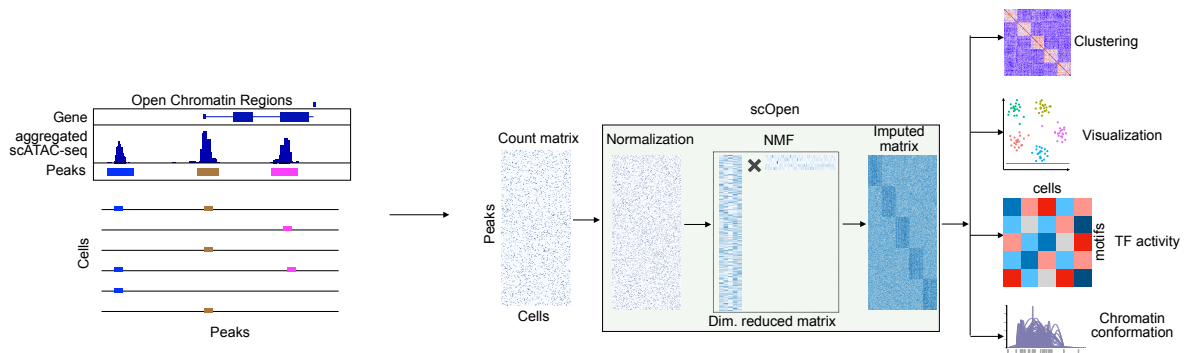
## 3.4 Implementation

We implemented our NMF-based strategy for single-cell ATAC-seq data imputation and dimensionality reduction as a Python command line tool. Our method is called scOpen (single-cell Open chromatin analysis via NMF modeling) and will be referenced as such throughout this thesis. Such a command line tool implements all the steps described in this chapter. For efficiency, scOpen uses

compressed sparse row matrix from SciPy (Virtanen et al., 2020) for data processing. The scOpen tool was first released in May 2021 and is available under GNU General Public License v3.0 (GPL v3). Figure 3.5 shows the workflow of the scOpen approach, and Table 3.1 summarizes the python package dependencies of scOpen.

The minimal input data required for scOpen is a peak by cell matrix and different formats for input data are accepted. For example, the matrix can be stored in a text file with dense format (i.e., each row represents a peak and each column represents a cell) or sparse format (each row contains the peak, cell and observed value). Moreover, scOpen can also take as input the outputs generated by Cell Ranger ATAC pipeline. The ability of supporting various input formats makes scOpen straightforward to use. The outputs of scOpen include: (i) an imputed matrix with the same dimensions as the input data; (ii) a text file representing the low-dimensional matrix. In addition, if the model selection options is enabled, a line plot describing the fitting error against the number of ranks will also be generated.

To ensure that scOpen displays interoperability with other popular Python-based frameworks, such as Scanpy (Wolf et al., 2018) and EpiScanpy (Danese et al., 2021), we have implemented an application programming interface (API) that allows for calling the functions in scOpen. Moreover, we provided a jupyter notebook (<https://github.com/CostaLab/scopen/blob/master/vignettes/epiScanpy.ipynb>) to illustrate how to operate scOpen with Scanpy and EpiScanpy. In addition, many popular tools for analyzing scATAC-seq data, including chromVAR (Schep et al., 2017), Signac (Stuart et al., 2020) and ArchR (Granja et al., 2021), are based on R. We therefore also provided a comprehensive online vignette ([https://github.com/CostaLab/scopen/blob/master/vignettes/signac\\_pbmc.Rmd](https://github.com/CostaLab/scopen/blob/master/vignettes/signac_pbmc.Rmd)) to demonstrate how to use scOpen under R environment.



**Figure 3.5: Workflow of scOpen.** scOpen receives as input a sparse peak by cell count matrix. After matrix binarization, scOpen performs TF-IDF transformation followed by NMF for dimension reduction and matrix imputation. The imputed or reduced matrix can then be given as input for scATAC-seq methods for clustering, visualization and interpretation of regulatory features

We have tested scOpen with Python 3.6-3.9 with Numpy 1.20.3, Scipy 1.6.3, H5py 3.2.1, PyTables 3.6.1, Matplotlib 3.4.2, Scikit-learn 0.24.2 and Knead 0.7.0. We used a local Linux Ubuntu 20.04 LTS x86 64-bit machine running with 8 Intel(R) Core(TM) i7-3770 CPU at 3.40GHz and 32 GB RAM. Moreover, we ran scOpen on an High Performance Computing (HPC) cluster mainly based on AMD

### 3.5. Discussion

Package	Version	Website
Numpy	>= 1.20.3	<a href="https://numpy.org/">https://numpy.org/</a>
Scipy	>= 1.6.3	<a href="https://www.scipy.org/">https://www.scipy.org/</a>
H5py	>=3.2.1	<a href="https://www.h5py.org/">https://www.h5py.org/</a>
Pandas	>= 1.2.4	<a href="https://pandas.pydata.org/">https://pandas.pydata.org/</a>
PyTables	>=3.6.1	<a href="https://www.pytables.org/">https://www.pytables.org/</a>
Matplotlib	>=3.4.2	<a href="https://matplotlib.org/">https://matplotlib.org/</a>
Scikit-learn	>=0.24.2	<a href="https://scikit-learn.org/stable/">https://scikit-learn.org/stable/</a>
Kneed	>=0.7.0	<a href="https://kneed.readthedocs.io/en/stable/">https://kneed.readthedocs.io/en/stable/</a>

**Table 3.1:** scOpen tool python package dependencies.

EPYC 7452 64-bit nodes at 2.35 GHz and 1024 GB RAM with CentOS Linux 8.

For more information about scOpen implementation, source code, tutorials and examples, please see:

<https://github.com/CostaLab/scopen>

## 3.5 Discussion

In this chapter, we presented our computational method, called scOpen, for scATAC-seq data analysis. We first described an approach for scATAC-seq data normalization (Section 3.2). Next, we introduced the NMF-based strategy for imputation and dimension reduction (Section 3.3) and implementation details (Section 3.4). The workflow of scOpen is showed in Figure 3.5 and a schematic overview is given by Algorithm 3.1. To summarize, our computational method applied new concepts to solve scATAC-seq imputation and dimension reduction problem:

- We introduced a novel scATAC-seq normalization framework that is able to correct potential technical effects, such as reads amplification and sequencing depth bias. Elements in the normalized matrix typically reflect the importance of a peak to a particular cell.
- We devised regularized NMF to factorize the original matrix into two smaller matrices (also known as factors) by considering the noise presented in the data. The resulting factors can be used as dimension reduces matrices for cells or peaks. Moreover, the multiplication of these two factors is regarded as an imputed or denoised matrix which is used for downstream analysis.
- We applied a computational algorithm to automatically estimate the number of components for NMF model. This method works by detecting the elbow point of reconstruction error against the rank of factors. We showed that in an example scATAC-seq dataset, this approach is able to identify the underlying dimensions, i.e., the number of cell types presented in the data.
- Although our method is implement by Python as a command line tool, we have provided a tutorial to demonstrate the interoperability between scOpen and other R-based tools, such as Signac (Stuart et al., 2020) and ArchR (Granja et al., 2021) (<https://github.com/>

CostaLab/scopen/blob/master/vignettes/signac\_pbmc.Rmd). We hope this will improve the availability of scOpen by allowing the users from R community to access scOpen without changing programming languages. Moreover, we also implemented an API that allows the other Python-based algorithms (e.g., EpiScanpy (Danese et al., 2021)) to directly use the results from scOpen (<https://github.com/CostaLab/scopen/blob/master/vignettes/epiScanpy.ipynb>).

From a methodological perspective, the core idea of our approach is the use of TF-IDF transformation and NMF. TF-IDF tends to adjust the weight of peaks to reflect the underlying importance of a peak to a cell. Moreover, the use of NMF for matrix imputation and dimension reduction is the favorable method of choice given that the TF-IDF transformed matrix is inherently non-negative. Comparing with the competing algorithms, including scImpute (Li and Li, 2018), SAVER (Huang et al., 2018), DCA (Eraslan et al., 2019), scBFA (Li and Quon, 2019), and SCALE (Xiong et al., 2019), our method made no assumptions about the data distribution in the matrix. Therefore, it is more robust when applying to a new dataset. Moreover, scOpen makes use of a linear approach, i.e., matrix factorization, to impute the matrix. However, MAGIC (Van Dijk et al., 2018) used manifold learning technique through a Gaussian kernel, which has been reported to generate many false positives in scRNA-seq context (Andrews and Hemberg, 2018). The regularization terms included in scOpen further prevented the model from overfitting. Regarding dimension reduction, cisTopic (González-Blas et al., 2019) used a similar topic modeling approach based on a Bayesian model and was reported to suffer from scalability problem. Table 3.2 summarizes the comparison between scOpen and other methods from methodological point of view.

Name	Technique	Imputation	Dimensionality Reduction	Distribution Assumption	Linearity	Regularization
scOpen	NMF	Yes	Yes	None	Linear	Yes
MAGIC	Manifold learning	Yes	No	None	Non-linear	No
scImpute	Probabilistic model	Yes	No	Gaussian	Non-linear	No
SAVER	Bayesian model	Yes	No	Poisson	Non-linear	No
DCA	Autoencoder	Yes	No	ZINB	Non-linear	No
scBFA	FA	Yes	No	Bernoulli	Non-linear	Yes
SCALE	VAE	Yes	Yes	GMM	Non-linear	No
cisTopic	LDA	Yes	Yes	Multinomial	Non-linear	No
LSI	SVD	No	Yes	None	Linear	No
SnapATAC	DM	No	Yes	None	Non-linear	No

**Table 3.2:** Methodological comparison between scOpen and other methods. NMF: non-negative matrix factorization, FA: factor analysis, LDA: Latent Dirichlet Allocation, VAE: variational autoencoder, SVD: singular value decomposition, DM: diffusion map, ZINB: zero-inflated negative binomial, GMM: Gaussian mixture model.



## Experiments

---

In the previous chapter, we introduced our computational approach, scOpen, for single-cell ATAC-seq imputation and dimensionality reduction. Here we present the experimental framework used in this thesis to validate our method. The framework is divided into two sections: technical validation and biological validation. In technical validation (Section 4.1), the major goal is to evaluate the performance of our approach and compare it to the competing methods. For biological validation (Section 4.2), we will apply scOpen to a novel scATAC-seq data generated from mouse kidney and test its power to dissect regulatory change in the development of fibrosis in the kidney. We finally close this chapter with a final discussion on our experimental workflow in Section 4.3.

### 4.1 Technical Validation

---

In this section, we present the experimental framework to technically validate our method. We first describe the scATAC-seq data used in this section (Section 4.1.1). Then, we outline the experiments to select the hyper-parameters in scOpen (Section 4.1.2). Next, We report the details of executing a number of computational approaches, including eight imputation methods, three dimensionality reduction methods, and three downstream analysis methods (Section 4.1.3). Finally, we describe the methodology used to evaluate the results produced by executing the computational methods (Section 4.1.4).

#### 4.1.1 Data

We here describe the process of generating scATAC-seq data for technical validation. We first introduce a customized pipeline for ATAC-seq data processing. Next, we detail our computational strategy to generate scATAC-seq simulation data. Finally, we describe the real-world scATAC-seq datasets that are publicly available. The simulation data is used to validate the parameter selection strategy in scOpen, and the benchmarking data is used to compare the performance of our approach against the competing methods.

#### ATAC-seq Processing Pipeline

We implemented a pipeline to preprocess ATAC-seq data from raw sequencing data to aligned files according to the description in Section 2.3.1. More formally, for a particular ATAC-seq library, we first converted the downloaded file to FastQ file using SRA toolkit (<http://ncbi.github.io/>

#### 4.1. Technical Validation

sra-tools/). Next, we trimmed adapter sequences and low-quality ends using Trim Galore (Martin, 2011). We mapped reads to the reference genome using Bowtie2 (Langmead and Salzberg, 2012) and removed the reads that were mapped to chrY, mitochondria, and unassembled "random" contigs. We also filtered out the duplicates with Picard (Institute, 2019) and only kept properly paired reads with alignment quality  $>30$ .

### Simulation Data

To generate simulated scATAC-seq data, we first obtained a bulk ATAC-seq with different cell types. For this, we downloaded the ATAC-seq of 13 human primary blood cell types from gene expression omnibus (GEO) with accession number GSE74912 (Corces et al., 2016). For each cell type, we processed the data using the pipeline as described above. Next, we called peaks using MACS2 (Zhang et al., 2008) and merged the peaks from all cell types to create a unique peaks list. We then created a peak cell-type matrix by offsetting +4 bp for forward strand and -5bp for reverse strand to represent the cleavage event center (Buenrostro et al., 2013; Li et al., 2019) and counting the number of reads start sites per cell type in each peak. This provides a peak by cell type matrix  $\mathbf{A}$ , where  $a_{ij}$  indicates the number of reads for peak  $i$  in cell type  $j$ .

We next used this bulk ATAC-seq counts matrix  $\mathbf{A}$  to simulate a scATAC-seq counts matrix  $\mathbf{X}$ . For this, we improved the simulation strategy proposed by Chen et al. (2019). Specifically, given  $m$  peaks and  $T$  cell types, to simulate a cell  $j$  for cell type  $t$ , we first sampled the total number of reads by:

$$N_j \sim NB(r, p) \quad (4.1)$$

where  $r$  and  $p$  parameterized a negative binomial distribution, and we estimated them using a real scATAC-seq dataset. Next, we introduced a parameter  $f$  to control the fraction of reads in peaks (FRiP) and computed the number of reads in peaks  $n_j$  by:

$$n_j = N_j \cdot f. \quad (4.2)$$

Next, we defined the rate at which the peak  $i$  is prevalent in bulk ATAC-seq data for cell type  $t$  as the ratio of reads observed in peak  $i$  over the total number of reads:

$$r_i^t = \frac{a_{it}}{\sum_{k=1}^m a_{kt}}. \quad (4.3)$$

Then, we estimated the probability of peak  $i$  being accessible in cell type  $t$  given the total number of reads  $n_j$  as follows:

$$p_i^t = r_i^t \cdot n_j \cdot (1 - q) + \left(\frac{1}{m}\right) \cdot n_j \cdot q \quad (4.4)$$

where  $q \in [0, 1]$  is a noise parameter. The probability  $p_i^t$  can be divided into the sum of two terms. The first term is the scaled ratio of reads for peak  $i$  from the bulk ATAC-seq data, and the second term represents a random distribution of  $n_j$  reads into  $m$  peaks. Intuitively, when  $q = 0$ , the simulated data is noiseless, and when  $q = 1$ ,  $p_i^t$  contains no cell-type-specific information. Finally, we obtained the



accessibility  $x_{ij} \in \{0, 1\}$  of cell  $j$  in peak  $i$  by sampling from a Bernoulli distribution:

$$x_{ij} \sim \text{Bernoulli}(p_i^t). \quad (4.5)$$

In total, we simulated 200 cells per cell type using the above process. We used noise  $q = 0.6$  and FRiP  $f = 0.3$ . Our approach differs from Chen et al. (2019) by sampling the number of reads per cell from a negative binomial distribution rather than using a fixed number (Equation 4.1). Moreover, we introduced the FRiP parameter (Equation 4.2). Algorithm 4.1 gives an overview of our simulation process.

---

**Algorithm 4.1:** Single-cell ATAC-seq data simulation

---

**Input** :  $\mathbf{A}$ , a  $m \times t$  matrix containing the number of reads for all peaks and cell types;  
 $m$ , the number of peaks;  
 $t$ , the number of cell types;  
**Parameter:**  $n$ , the number of simulated cells for each cell type;  
 $q$ , the noise parameter;  
 $\text{FRiP}$ , the fraction of reads in peaks;  
**Output** :  $\mathbf{X}$ , a  $m \times n$  accessibility matrix;

```

1 for  $t \leftarrow 1$  to  $T$  by 1 ;                                // Go over all cell types
2 do
3   for  $j \leftarrow 1$  to  $n$  by 1 ;                            // Go over all cells
4   do
5     for  $i \leftarrow 1$  to  $m$  by 1 ;                          // Go over all peaks
6     do
7       Sample the total number of reads  $N_j$  using Equation 4.1;
8       Calculate the number of reads in peaks  $n_j$  using Equation 4.2;
9       Compute the rate  $r_i^t$  for peak  $i$  and cell type  $t$  using Equation 4.3;
10      Estimate the probability  $p_i^t$  for peak  $i$  and cell type  $t$  using Equation 4.4;
11      Sample the observation  $x_{ij}$  using Equation 4.5;
12    end
13  end
14 end
15 return  $\mathbf{X}$ 

```

---

## Benchmarking Data

The benchmarking data is composed of four real-world scATAC-seq datasets generated by using either plate-based protocol or droplet-based protocol. We selected these datasets due to the presence of external labels, which were defined independently of the scATAC-seq at hand. We used the labels as ground truth for evaluation. See Table 4.1 for complete statistics associated with these datasets.

**Cell line** This dataset was obtained by combining scATAC-seq data from six cell types, namely BJ, H1-ESC, K562, GM12878, TF1, and HL-60 from Buenrostro et al. (2015b). The data was generated using the plate-based scATAC-seq protocol. For every single cell, we downloaded the

#### 4.1. Technical Validation

sequencing data from GEO with accession number GSE65360 and processed the data using the pipeline. For quality control, we only kept cells with more than 500 unique fragments. Then, we created a pseudo-bulk ATAC-seq library by merging all cells. We next called peaks using MACS2 (Zhang et al., 2008) and extended the peaks by  $\pm 250$ bp from the summits as in Buenrostro et al. (2013). After processing and filtering, we obtained 1,224 cells and 125,647 peaks. We then constructed a peak by cell count matrix and used it as input for evaluation. The cell types included in this data are quite different from each other, and they should be easily separated by clustering, meaning that this dataset mainly serves as a baseline for benchmarking. We will refer to this dataset as *Cell line*.

**Hematopoiesis** This dataset includes single-cell chromatin accessibility profiles across eight immunophenotypically defined human hematopoietic cell types: hematopoietic stem cells (HSC), multipotent progenitors (MPP), lymphoid-primed multi-potential progenitors (LMPP), common myeloid progenitors (CMP), common lymphoid progenitors (CLP), granulocyte-macrophage progenitors (GMP), megakaryocyte-erythroid progenitors (MEP) and plasmacytoid dendritic cells (pDC) (Buenrostro et al., 2018). The data was generated using the plate-based scATAC-seq protocol. The cell labels were defined by using cell surface markers. We processed the data as the same as we did for the dataset *Hematopoiesis*. Finally, we obtained a count matrix with 2,210 cells and 109,418 peaks. It is worth pointing out that these cells assemble a continuous process about human hematopoietic differentiation. Therefore, it is relatively hard to cluster the cells. We will refer to this dataset as *Hematopoiesis*.

**T cells** This dataset is composed of single-cell chromatin accessibility data from human T cell subpopulations, i.e., Jurkat T cells, memory T cells, naive T cells, and Th17 T cells. The sequencing data was obtained from GSE107816 (Satpathy et al., 2018) and was processed as described above. Labels were provided in Satpathy et al. (2018) by comparing the profiles to bulk ATAC-seq of corresponding T-cell subpopulations. We finally obtained 765 cells and 49,344 peaks. Although this dataset contains much fewer cells, it represents a harder problem for clustering compared with *Cell line* and *Hematopoiesis*, given the highly similar chromatin accessibility profile of different T cell subpopulations. We will refer to this dataset as *T cells*.

**PBMC** To test the scalability of imputation and dimensionality reduction methods, we also included a multiome peripheral blood mononuclear cells (PBMC) dataset that contains about 10,000 cells with 14 cell types. The data was generated using the Chromium Single Cell Multiome ATAC + Gene Expression assay which simultaneously profiles the epigenomic landscape and gene expression in the same single nuclei. We downloaded the data as a Seurat object from [https://raw.githubusercontent.com/bioFAM/MOFA2\\_tutorials/master/R\\_tutorials/10x\\_scRNA\\_scATAC.html](https://raw.githubusercontent.com/bioFAM/MOFA2_tutorials/master/R_tutorials/10x_scRNA_scATAC.html). This object contained the count matrix which was used as input in the following analysis. For evaluation, we used the cell types annotated by the 10X Genomics R&D team based on the scRNA-seq modality alone. We will refer to this dataset as *PBMC*.

Dataset	Number of cells	Number of features	Fraction of non-zeros	Number of reads per cell	FRiP	Number of total reads
Cell lines	1,224	125,647	0.036	41,467.80	0.248	50,756,587
Hematopoiesis	2,210	109,418	0.039	34,656.15	0.272	76,590,091
T cells	765	49,344	0.033	14,963.39	0.418	11,446,993
PBMC	10,032	106,935	0.067	13,486	0.714	457,001,034

**Table 4.1: Statistics of the benchmarking datasets used in this thesis.** For each dataset, the number of detected cells, the number of regions (peaks), the fraction of non-zero entries, the average number of reads per cell, the fraction of reads in peaks (FRiP), and the total number of valid reads are showed.

#### 4.1.2 scOpen Parameter Selection

As described in Section 3.3.4, there are two hyper-parameter in scOpen, i.e., regularization parameter  $\lambda$  and the number of components  $k$ . We have introduced a computational approach to select the number of components in the model. However, there is no such a way to automatically determine the  $\lambda$ . Therefore, we here used the simulation data to perform model selection. More specifically, we evaluated the model performance by using a number of  $\lambda$  values with two metrics. Furthermore, we also verified our strategy to determine the parameter  $k$ .

#### 4.1.3 Execution of Computational Methods

In this section, we present the full details of the parameterization and execution of computational methods that are used evaluated in this thesis.

#### Competing Imputation Methods

We first describe the execution of computational imputation methods that were described in Section 2.4.2. In addition, we also included a PCA-based method (termed here as imputePCA) as a control for comparison.

**MAGIC** We installed the MAGIC R and Python package (v3.0.0) according to the tutorial <https://github.com/KrishnaswamyLab/MAGIC>. We applied it to the count matrix with the default setting. Prior to MAGIC, the count matrix was first normalized by library size and then root squared, as suggested by the authors (Van Dijk et al., 2018).

**SAVER** We obtained SAVER (v1.1.2) from <https://github.com/mohuangx/SAVER> and ran it on the count matrix. SAVER has two main steps: the first is the prediction step, and the second is a shrinkage step. We followed the tutorial <https://mohuangx.github.io/SAVER/articles/saver-tutorial.html> to perform denoising.

**scImpute** We downloaded scImpute (v0.0.8) from <https://github.com/Vivianstats/scImpute> and executed it using the default setting. It is worth pointing out that scImpute requires the number of cell sub-populations as an input parameter to determine the candidate

#### 4.1. Technical Validation

neighbors of each cell. For this, we used the true cluster number from each benchmarking dataset.

**DCA** We installed DCA (v0.3.1) from <https://github.com/theislab/dca> and ran the autoencoder from the command line with the default setting. For evaluation, we used the output, which represents the mean parameters of the ZINB distribution.

**cisTopic-impute** We downloaded cisTopic (v2.1.0) from <https://github.com/aertslab/cisTopic> and ran it with different numbers of topics (from 5 to 50). The optimal number of topics was selected based on the highest log-likelihood, as suggested by González-Blas et al. (2019). We then multiplied the topic-cell and the region-topic distributions to obtain the predictive distribution González-Blas et al. (2019), which describes the probability of each region in each cell and is used as the imputed matrix for clustering and visualization. We call this method *cisTopic-impute*.

**scBFA** scBFA is a detection-based model to remove technical variation for both scRNA-seq and scATAC-seq by analyzing feature detection patterns alone and ignoring feature quantification measurements Li and Quon (2019). We obtained scBFA (v1.0) from <https://github.com/quon-titative-biology/scBFA> and ran it on the raw count matrix using default parameters.

**SCALE** SCALE combines the variational auto-encoder (VAE) and the Gaussian Mixture Model (GMM) to model the distribution of high-dimensional sparse scATAC-seq data Xiong et al. (2019). We downloaded SCALE (v1.1.0) from <https://github.com/jsxlei/SCALE> and ran it with the default setting. We used option *-impute* to get the imputed data.

**imputePCA** We also included principal component methods (termed here as imputePCA) on incomplete data sets as a control for comparison. This method is based on an interactive and regularized PCA algorithm to predict missing entries, which are considered as latent variables (Josse and Husson, 2016). We installed R package missMDA (v1.18) from <https://cran.r-project.org/web/packages/missMDA/index.html> and performed imputation with function *imputePCA* with default settings. All zero entries were considered as missing data.

#### Dimensionality Reduction Methods

Besides imputation, scOpen also provides a dimensions reduced matrix for input scATAC-seq data. Here, we also compared scOpen with the state-of-the-art scATAC-seq dimension reduction methods: cisTopic (González-Blas et al., 2019), SnapATAC (Fang et al., 2021), and latent semantic indexing (LSI) (termed here as Cusanovich2018 (Chen et al., 2019)), as described in Section 2.4.3. We applied these methods to obtain a low-dimensional matrix for each benchmarking dataset as detailed below.

**cisTopic** We executed cisTopic as described above and used the topic-cell distribution as dimensions reduced matrix.

**SnapATAC** We installed SnapATAC (v2.0) from <https://github.com/r3fang/SnapATAC>. Instead of using the count matrix, SnapATAC takes as input the fragments and produces a low-dimensional matrix as output. Moreover, it specifically works on the snap (Single-Nucleus Accessibility Profiles) file, which is a hierarchically structured hdf5 file. In order to generate such input files as required by SnapATAC, for each of the plate-based datasets (including *Cell line*, *Hematopoiesis*, and *T cells*), we first converted the BAM file to a BED file which contains the fragments for all cells using the function `readBamFileAsGRanges` from R package `chromstaR` (Taudt et al., 2016). For *PBMC*, we downloaded the file from <https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets>. Next, we installed SnapTools (v1.2.3) from <https://github.com/r3fang/SnapTools> to process the fragments to generate the snap files. We then followed the tutorial [https://github.com/r3fang/SnapATAC/blob/master/examples/10X\\_brain\\_5k/README.md](https://github.com/r3fang/SnapATAC/blob/master/examples/10X_brain_5k/README.md) to compute diffusion maps for dimensionality reduction.

**Cusanovich2018** For each dataset, we first segmented the whole genome into 5kb windows and then scored each cell for any insertions in these windows. This generated a large, sparse, and binary matrix of 5kb windows by cells. Based on this matrix, we retained the top 20,000 most commonly used sites. Then, we normalized and re-scaled the matrix using TF-IDF transformation using the function `RunTFIDF` from R package `Signac` (Stuart et al., 2020). Finally, we performed singular value decomposition (SVD) to generate a PCs-by-cells low dimensional matrix using the function `RunSVD`.

## Downstream Analysis Methods

We also tested whether the scOpen imputed matrix benefits the downstream analysis of scATAC-seq. The intuition is that if we improve the count matrix by imputation, we should be able to improve downstream analysis. For this, we selected three state-of-the-art methods for scATAC-seq analysis, namely, scABC (Zamanighomi et al., 2018), chromVAR (Schep et al., 2017), and Cicero (Pliner et al., 2018). We applied these methods using either scOpen imputed matrix or raw scATAC-seq count matrix as input.

**Cicero** Cicero is a method that predicts co-accessible pairs of DNA elements using single-cell chromatin accessibility data (Pliner et al., 2018). Moreover, Cicero provides a gene activity score for each cell and gene by assessing the overall accessibility of a promoter and its associated distal sites. We installed Cicero (v1.3.0) from <https://cole-trapnell-lab.github.io/cicero-release/>. For each benchmarking dataset, we followed the document provided by <https://cole-trapnell-lab.github.io/cicero-release/docs/> to generate a gene activity matrix which was used for clustering and visualization of scATAC-seq data.

**chromVAR** chromVAR is an R package for analyzing sparse chromatin-accessibility data by measuring the gain or loss of chromatin accessibility within sets of genomic features, as regions

## 4.1. Technical Validation

with sequence predicted transcription factor (TF) binding sites (Schep et al., 2017). We obtained chromVAR (v1.14.0) from <https://github.com/GreenleafLab/chromVAR>. For each dataset, we first computed the GC content for the peaks using the function *addGCBias* and then detected TF binding sites using the function *matchMotifs*. The motifs were obtained from the JASPAR database with version 2020 (Fornes et al., 2020). Finally, we estimated the deviations in chromatin accessibility across the TF binding sites using the function *computeDeviations*. The deviation scores were used to cluster the cells using 1 - Pearson correlation as distance.

**scABC** scABC is an unsupervised clustering algorithm for single-cell epigenomic data (Zamanighomi et al., 2018). The algorithm can be broken down into three steps. First, the weighted *K*-medoids clustering method is used to obtain the initial cluster assignment for every single cell. Second, the reads of cells within a cluster are summed and a number of peaks with the highest read counts are obtained to build a landmark for each cluster. Third, the cells are re-clustered by assigning each cell to the landmark with the highest correlation using the union of all landmark peaks. We downloaded scABC (v0.1) from <https://github.com/SUwonglab/scABC> and executed it by following the tutorial <https://github.com/SUwonglab/scABC/blob/master/vignettes/ClusteringWithCountsMatrix.Rmd>. Finally, we evaluated the clustering results generated by scABC.

### 4.1.4 Evaluation of Computational Methods

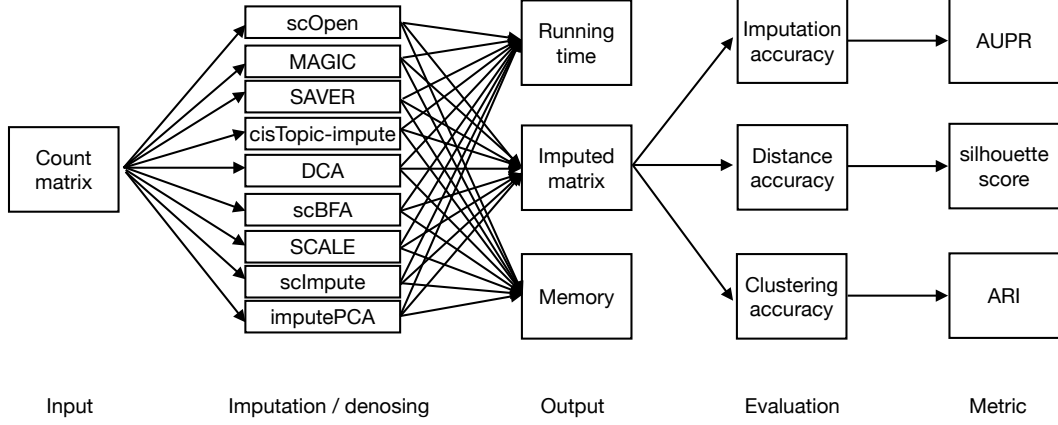
In this section, we present the methodology used to evaluate the results generated by execution of the imputation, dimensionality reduction, and downstream analysis methods as previously described.

#### Imputation Methods

We used several metrics to benchmark the imputation methods, i.e., memory and running time requirements, imputation accuracy, distance accuracy, and clustering accuracy. These metrics evaluated the results from different perspectives. Figure 4.1 depicts the overview framework of the evaluation of the imputation methods.

**Requirements of Memory and Running Time** To compare the memory and running time requirements of each imputation method, we ran all of them on a dedicated HPC node with the same computation resources quota, i.e., 180GB memory, 120 hours, and 4 CPUs. For DCA and SCALE, two deep learning-based methods, we used GPU with 16GB memory. We measured the max memory usage during the running of a method and recorded the total running time for each method.

**Imputation Accuracy** This was used to test if the imputation methods can improve the detection of true open chromatin (OC) regions for every single cell. In order to perform this evaluation, we first defined the ground truth labels for each cell. For this, we created a bulk ATAC-seq profile for each cell type by aggregating the data from all cells within that cell type with SAMtools (Li



**Figure 4.1: Experimental design for evaluation of the imputation methods.** We applied the imputation methods on each of the benchmarking datasets to generate an imputed matrix. Next, we evaluated the results based on imputation accuracy (measured using AUPR), distance accuracy (measured using silhouette score), and clustering accuracy (measured using ARI).

et al., 2009). Next, we performed peak calling using MACS2 (Zhang et al., 2008) to identify cell-type-specific OC regions. These OC regions present in a particular cell type were considered as positive labels, and OC regions not present in that cell type as negative labels. Then, for every single cell, we used the labels from the corresponding cell type as ground truth.

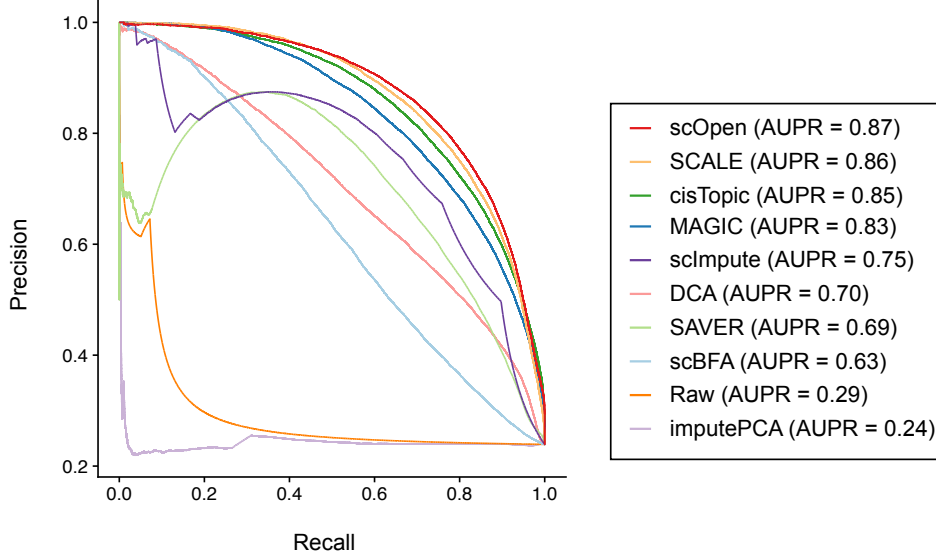
Next, we evaluated the imputation results for every single cell against the true labels. Because the labels are highly imbalanced, i.e., there are much more negatives than positives (Appendix Table A.1), we chose to use Precision-Recall (PR) metric to evaluate the prediction quality (Davis and Goadrich, 2006). More formally, for a specific cell, we defined the precision ( $P$ ) and recall ( $R$ ) of the prediction given a certain threshold as follows:

$$\begin{aligned}
 P &= \frac{TP}{TP + FP} \\
 R &= \frac{TP}{TP + FN}
 \end{aligned}
 \tag{4.6}$$

where  $TP$  represents the number of true positives,  $FP$  represents the number of false positives, and  $FN$  means the number of false negatives. Intuitively, a model with high precision and low recall means that it predicts very few samples as positive, but most of the predictions are correct. However, a model with high recall and low precision is just the opposite. We obtained a PR curve and computed the area under the curve (termed as AUPR) as the metric for every single cell using the PRROC package (Grau et al., 2015). Figure 4.2 gives an example of the PR curve for a specific cell from benchmarking dataset *Cell line*.

**Distance Accuracy** We also measured how similar a cell is to the cells with the same label compared to other cell types after imputation. The logic is that if the imputation is working, the cells from the same cell type should tend to show a high cohesion and vice versa. For this evaluation, we calculated a silhouette score (Rousseeuw, 1987) for each cell and imputation method. More specifically, given a cell  $i$  from cell type  $C_k$ , we first computed the average distance between  $i$

#### 4.1. Technical Validation



**Figure 4.2: An example of the Precision-Recall curves.** This figure shows an example of the Precision-Recall curves that compares the imputation methods for a single cell in terms of peak recovery. The colors refer to imputation methods, and AUPR represents the area under PR curve.

and the other cells from the same cell type as following:

$$a(i) = \frac{1}{|C_k - 1|} \sum_{j \in C_k, j \neq i} d(i, j) \quad (4.7)$$

where  $d(i, j)$  represents the distance between cell  $i$  and  $j$ . In this evaluation, we estimated the distance as  $1 - \text{Pearson correlation}$ . Next, we computed the smallest mean distance between  $i$  and all cells from other cell types as following:

$$b(i) = \min_{k'} \frac{1}{|C_{k'}|} \sum_{j \in C_{k'}} d(i, j). \quad (4.8)$$

This represents the distance between  $i$  and its next nearest cluster centroid (Rousseeuw, 1987). Finally, we calculated the silhouette score for cell  $i$  according to:

$$s(i) = \begin{cases} \frac{b(i) - a(i)}{\max(a(i), b(i))}, & \text{if } |C_k| > 1 \\ 0, & \text{otherwise.} \end{cases} \quad (4.9)$$

It is clear that  $-1 \leq s(i) \leq 1$ , and a higher silhouette score indicates a higher similarity of a cell to the cells of the same cell type than the cells from other cell types.

**Clustering Accuracy** Clustering is one of the core components for scATAC-seq analysis, as it forms the basis for various downstream analyses. To evaluate the performance of clustering after imputation, we applied PCA (50 PCs) for the imputed matrix to first generate a low-dimensional representation. Next, we clustered the cells using  $1 - \text{Pearson correlation}$  as distance. To avoid information leakage and enable a statistical comparison, we included two clustering algorithms,



i.e., k-medoids and hierarchical clustering. Besides PCA, we also used t-SNE (van der Maaten and Hinton, 2008) embedding as input and euclidean as distance, given that this approach is also explored by cisTopic (González-Blas et al., 2019). Furthermore, we also tested the different numbers of clusters, e.g.  $k$  and  $k + 1$ , where  $k$  is the true number of clusters for the corresponding dataset.

For comparison, we used the adjusted rand index (ARI) (Hubert and Arabie, 1985) to evaluate the clustering results with the labels from each of the benchmarking datasets. The adjusted rand index measures similarity between two data clustering results by correcting the chance of grouping elements. More specifically, given a dataset  $D$  with  $n$  cells, two partitions  $U = \{U_1, U_2, \dots, U_r\}$  and  $V = \{V_1, V_2, \dots, V_s\}$  representing different clustering results, the number of common cells for each cluster  $i$  and  $j$  can be written as:

$$c_{ij} = |U_i \cap V_j| \quad (4.10)$$

where  $i \in \{1, 2, \dots, r\}$  and  $j \in \{1, 2, \dots, s\}$ . The ARI can be calculated as following:

$$ARI = \frac{\sum_{ij} \binom{c_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}} \quad (4.11)$$

where  $a_i = \sum_{j=1}^s c_{ij}$  and  $b_j = \sum_{i=1}^r c_{ij}$ , respectively. The ARI has a maximum value 1 and an expected value 0, with 1 indicating that the data clustering is the exact same and 0 indicating that the two data clustering agree randomly.

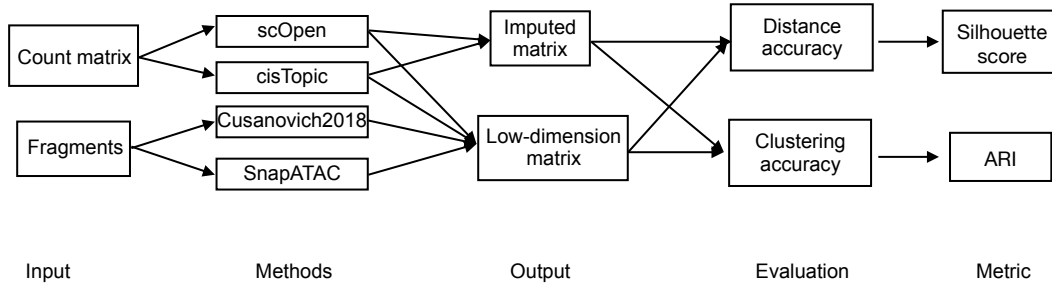
## Dimensionality Reduction Methods

To evaluate the performance of the selected dimensionality reduction methods (Section 2.4.3), we applied them to each of the benchmarking datasets to obtain a dimensions reduced matrix. We measured the distance accuracy using silhouette score as previously described (Equation 4.9). To achieve a fair comparison about clustering accuracy, we used a density-based clustering approach for scOpen, cisTopic, and Cusanovich2018, and a graph-based clustering method for SnapATAC as proposed in the original papers. We also evaluated the use of both reduced and imputed matrices for scOpen and cisTopic, as these methods provide both types of representations. Figure 4.3 depicts the overview framework of evaluation of the dimension reduction methods.

## Downstream Analysis Methods

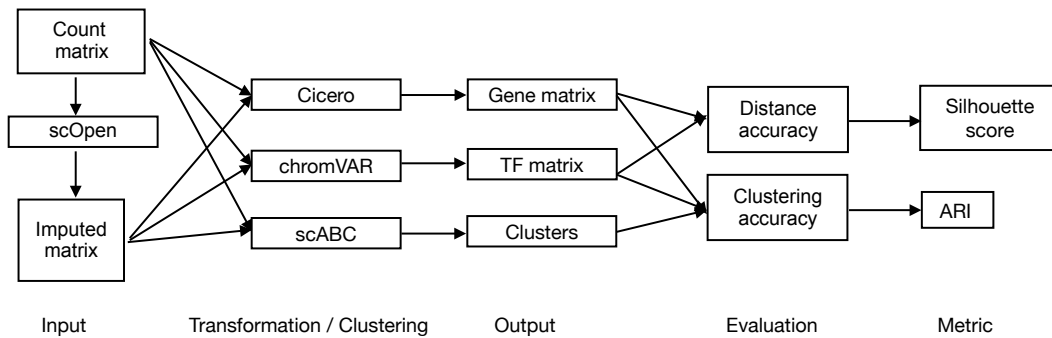
To test if scOpen can improve the performance of scATAC-seq downstream analyses (e.g., gene expression prediction by Cicero (Pliner et al., 2018), motif analysis by chromVAR (Schep et al., 2017), and clustering by scABC (Zamanighomi et al., 2018)), we applied these methods to either scATAC raw count matrix or scOpen imputed matrix for each of the benchmarking datasets as previously described. Cicero and chromVAR transformed the peak by cell matrix to a gene by cell matrix and a TF by cell matrix, respectively. The outputs were evaluated based on distance accuracy as measured by silhouette score and clustering accuracy as measured by ARI. For scABC, we directly used the clus-

#### 4.1. Technical Validation



**Figure 4.3: Experimental design for evaluation of the dimensionality reduction methods.** We applied the dimensionality reduction methods on each of the benchmarking datasets to generate a low-dimensional matrix. Next, we evaluated the results based on distance accuracy (measured using silhouette score) and clustering accuracy (measured using ARI).

tering results and calculated the ARI. Figure 4.4 shows an overview the framework for evaluation of the downstream analysis methods.

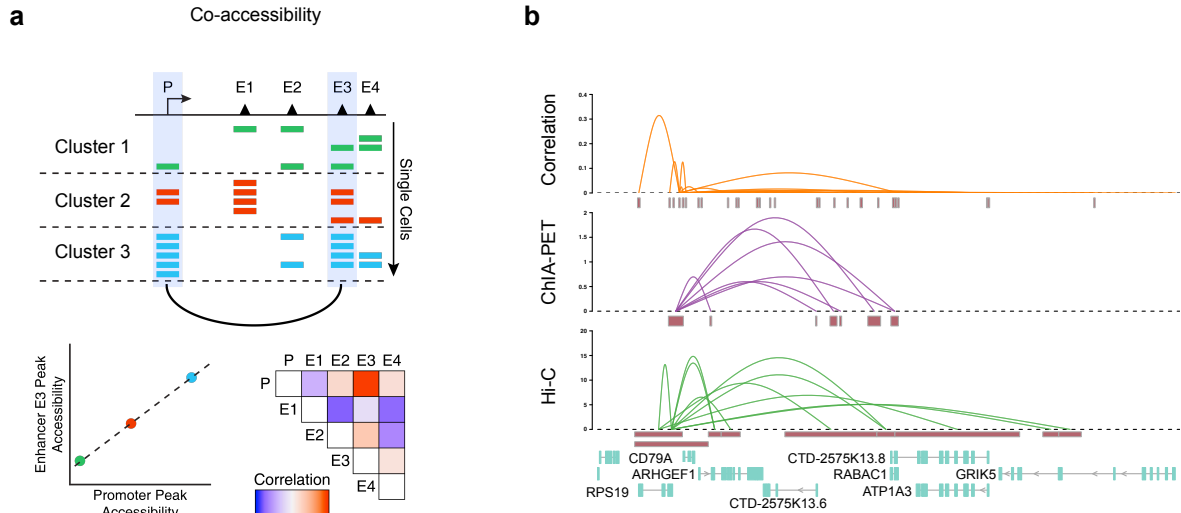


**Figure 4.4: Experimental design for evaluation of the downstream analysis methods.** We applied the downstream analysis method on each of the benchmarking datasets using either raw count matrix or scOpen imputed matrix as input. The outputs were evaluated based on different metrics.

#### Co-accessibility Analysis

We also tested whether or not the imputed matrix improves the prediction of co-accessible pairs, given that the sparsity has been reduced. For this, we downloaded the scATAC-seq matrix of GM12878 cells from GEO with the accession number GSM2970932 and applied the imputation methods to generate an imputed matrix as described above (Section 4.1.3). Next, we predicted co-accessible peaks using Cicero based on either raw count or imputed matrix of GM12878. Figure 4.5a depicts the theoretical principle of co-accessibility analysis. For evaluation, we used conformation data as true labels, given that co-accessible peaks have been reported to have a high agreement with previously observed chromosome compartments (Buenrostro et al., 2015b; Kalhor et al., 2012). We downloaded promoter-capture (PC) Hi-C data of GM12878 from GEO (GSE81503), which used ChICAGO (Cairns et al., 2016) score as a physical proximity indicator. We also downloaded ChIA-PET data of GM12878 from GEO (GSM1872887), which used the frequency of each interaction PET cluster to represent how strong the interaction is. We considered all obtained links, as provided by these data sets, as

true interactions as in (Pliner et al., 2018). Next, we replicated the evaluation analysis performed in (Pliner et al., 2018) and contrasted the results of Cicero with raw or matrices obtained after imputation. Figure 4.5b compares the predicted links by Cicero based on raw matrix with scores based on ChIA-PET or Hi-C protocol. Next, we used the built-in function *compare\_connections* of Cicero to define the true labels for predicted co-accessibility links. Using the correlation as prediction, we finally computed the AUPR values with the function *pr.curve* from the R package PRROC (Grau et al., 2015). To investigate the performance of each method against the number of cells, we also randomly down-sampled the data to 50% and 25%, and repeated the above analysis.



**Figure 4.5: Co-accessibility prediction and evaluation.** **a**, Schematic workflow for predicting co-accessible peaks from a scATAC-seq matrix. The cells are first clustered and aggregated. Next, correlations between peaks are calculated. *Source: Granja et al. (2021)* (modified to fit thesis format and/or clarify key points). **b**, Visualization of co-accessibility scores (y-axis) of Cicero predicted with raw matrix contrasted with scores based on RNA pol-II ChIA-PET (purple) and promoter capture Hi-C (green) around the CD79A locus (x-axis). For ChIA-PET, the log-transformed frequencies of each interaction PET cluster represent co-accessibility scores, while the negative log-transformed p-values from the CHiCAGO software indicate Hi-C scores.

#### 4.1.5 Statistical Methods

For comparisons involving multiple methods and datasets, we used the non-parametric Friedman test with the Nemenyi post-hoc test (Demšar, 2006). The Friedman test (Friedman, 1937, 1940) was used to compare the average ranks of the methods across all datasets. The null hypothesis is that all the methods are equivalent, and their ranks are equal. We used the function *friedmanTest* from the R package PMCMRplus to perform the Friedman rank sum test. If the null hypothesis was rejected, meaning that at least one method is significantly different from other methods. We then performed the Nemenyi post-hoc test (Nemenyi, 1963) to compare all pairs. Therefore, to compare  $k$  methods, a total of  $k(k-1)/2$  hypotheses were tested. This was done using the function *frdAllPairsNemenyiTest*

## 4.2. Biological Validation

from R package PMCMRplus.

To compare the differences between the distributions of two groups or more groups in a pair-wise manner, we used the non-parametric Mann–Whitney–Wilcoxon rank-sum test (Mann and Whitney, 1947). The null hypothesis is that the distributions of both groups are equal. All test p-values were calculated based on confidence levels of 0.95 and the continuity correction was always applied in the normal approximation for the p-value. We used the function *wilcox.test* from the R programming language version 4.0.3 implementation to perform such a test. Moreover, in case of multiple testing, we used the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to correct the p-values. This method is able to control the false discovery rate (FDR) in a practical and powerful manner. We used the function *p.adjust* from the R programming language version 4.0.3 implementation to carry out such correction.

## 4.2 Biological Validation

---

In this section, we present the experimental details to biologically validate our method. In particular, we applied our computational method, scOpen, to a novel scATAC-seq dataset and demonstrated its power to capture regulatory dynamics in a complex fibrosis system. The work presented in this section results from a collaboration with the group of Prof. Rafael Kramann in the Institute of Experimental Medicine and Systems Biology at the RWTH Aachen University Medical School. The biological experiments were carried out by Christoph Kuppe, Susanne Ziegler, Nazanin Kabgani, and Sylvia Menzel as specifically described below. All the computational analyses were performed by me.

### 4.2.1 Applying scOpen to scATAC-seq Data from Complex Disease

#### Generation of scATAC-seq Data from Mouse Kidneys

We performed unilateral ureter obstruction (UUO) as previously described (Kramann et al., 2015). Shortly, the left ureter was tied off at the level of the lower pole with two 7.0 ties (Ethicon) after flank incision. One C57BL/6 male mouse (age 8 weeks) was sacrificed on day 0 (sham), day 2, and 10 after the surgery. Kidneys were snap-frozen immediately after sacrifice. Pdgfrb-BAC-eGFP reporter mice (for staining experiments, age 6-10 weeks, C57BL/6) were developed by N. Heintz (The Rockefeller University) for the GENSAT project. Genotyping of all mice was performed by PCR. Mice were housed under specific pathogen-free conditions at the University Clinic Aachen. Pdgfrb-BAC-eGFP were sacrificed on day 10 after the surgery. All animal experiment protocols were approved by the LANUV-NRW, Düsseldorf, Germany. All animal experiments were carried out in accordance with their guidelines.

Next, we performed nuclei isolation as recommended by 10X Genomics (demonstrated protocol CG000169). The nuclei concentration was verified using stained nuclei in a Neubauer chamber with trypan-blue targeting a concentration of 10,000 nuclei. Tn5 incubation and library preparation was done by following the 10X scATAC protocol. After checking the quality using Agilent BioAnalyzer, we pooled the libraries. Finally, we performed sequencing on a NextSeq in 2x75bps paired-end run

with three runs of the NextSeq 500/550 High Output Kit v2.5 Kit (Illumina), resulting in more than 600 million reads. Sylvia Menzel took care of the animal experiments (breeding, operation etc) and Christoph Kuppe performed the UUO surgery, nuclei isolation, and the sequencing.

### Computational Analysis of scATAC-seq Data with scOpen

We used Cell-Ranger ATAC (v1.1.0) pipeline to perform low-level data processing. We first demultiplexed raw base call files using *cellranger-atac mkfastq* with the default setting to generate FASTQ files for each flow-cell. Next, we applied *cellranger-atac count* to perform read trimming, filtering, and alignment. We then estimated the transcription start site (TSS) enrichment score using the obtained fragment files and filtered out low-quality cells using a TSS score of 8 and a number of unique fragments of 1,000 as thresholds. The remaining barcodes were considered as valid cells for further analysis. We next performed peak calling using MACS2 for each sample and merged the peaks to generate a union peak set, which was used to create a peak by cell matrix.

We next applied scOpen to generate a low-dimensional representation of the cells. For comparison, we also executed the competing methods (i.e., cisTopic, SnapATAC, and LSI (termed here Cusanovich2018)) as previously described. We then used Harmony Korsunsky et al. (2019) to integrate the scATAC-seq profiles from different time points (day 0, day 2, and day 10) using either Cusanovich2018, cisTopic, scOpen, or SnapATAC dimension reduced matrix as input. Specifically, we created a Seurat object for each of the low-dimension matrices and ran the Harmony algorithm with the function *RunHarmony*. Next, we used k-medoids to cluster the cells by taking batch-corrected low-dimension matrix as input. The number of clusters was set to 17, given that the single-nucleus RNA-seq that we used as a reference for annotation identified 17 unique cell types (See below).

To evaluate and annotate the clusters obtained from data integration, we downloaded a publicly available snRNA-seq dataset of the same fibrosis model (GSE119531) and performed label transfer using Seurat3 (Stuart et al., 2019). This dataset contains 6,147 single-nucleus transcriptomes with 17 unique cell types (Wu et al., 2019). For label transferring, we used the gene activity score matrix estimated by ArchR and transferred the cell types from the snRNA-seq dataset to the integrated scATAC-seq dataset by using the functions *FindTransferAnchors* and *TransferData* in Seurat3 (Stuart et al., 2019). For benchmarking purposes, the predicted labels were used as the true labels to compute ARI for evaluation of the clustering results and silhouette score for evaluation of the distance accuracy after using different dimension reduction methods as input for data integration. We also performed the same analysis for each sample separately and computed the metrics.

### Cell Annotation

For the biological interpretation, we estimated doublet scores using ArchR (Granja et al., 2021) and removed cells with a doublet score above 2.5. Next, we named the cluster by assigning the label with the highest proportion of cells to the cluster and checking marker genes. In total, we recovered 16 unique cell types from the 17 labels, as two clusters (2 and 17) were annotated as TAL cells. Specifically, we denoted clusters 6, 1, 3 as proximal tubule (PT) S1, S2, and S3 cells. We annotated cluster 2 as thick ascending limb (TAL), cluster 5 as distal convoluted tubule (DCT), cluster 7 as

## 4.2. Biological Validation

collecting duct-principal cell (CD-PC), cluster 8 as endothelial cell (EC), cluster 9 as connecting tubule (CNT), cluster 10 as intercalated cell (IC), cluster 11 as fibroblast, cluster 12 as descending limb + thin ascending limb (DL & TAL), cluster 13 as macrophage (MAC), cluster 16 as podocytes (Pod). Cluster 14 was identified as injured PT, which was not described in Wu et al. (2019), given the increased accessibility of marker *Vcam1* and *Havcr1*. We also renamed the cells of cluster 15, which were labeled as *Mac2* in (Wu et al., 2019), as lymphoid cells given that these cells express B and T cell markers *Ltb* and *Cd1d*, but not macrophage markers *C1qa* and *C1qb*. Finally, cluster 4 was removed based on the doublet analysis. Prof. Rafael Kramann and Christoph Kuppe supported the cell annotation.

### Estimation of Cell-type-specific TF Activity

We adapted the differential TF activity analysis from HINT-ATAC (Li et al., 2019) for scATAC-seq. In short, we created pseudo-bulk ATAC-seq libraries by combining reads of cells for each cell type and performed footprinting with HINT-ATAC. Next, we predicted TF binding sites by motif analysis (FDR = 0.0001) inside footprint sequences using the RGT toolkit (v0.12.3). Motifs were obtained from JASPAR Version 2020 (Fornes et al., 2020). We measured the average digestion profiles around all binding sites of a given TF for each pseudo bulk ATAC-seq library. We then used the protection score (Li et al., 2019), which measures the cell-specific activity of a factor by considering the number of digestion events around the binding sites and depth of the footprint. Higher protection scores indicated a higher activity (binding) of that factor. Finally, we only considered TFs with more than 1,000 binding sites and variance in activity score higher than 0.3. We also performed smoothing for visualization of average footprint profiles. In short, we performed a trimmed mean smoothing (5 bps window) and ignored cleavage values in the top 97.5% quantile for each average profile.

## 4.2.2 Characterizing Gene Regulation During Myofibroblast Differentiation

### Computational Identification of Key TFs for Myofibroblast Differentiation

We performed sub-clustering of fibroblast cells on batch-corrected low-dimension scOpen matrix. In total, we obtained three clusters which were annotated as pericyte (cluster 1), myofibroblast (cluster 2), and Scara5+ fibroblast (cluster 3) using known marker genes, respectively. For visualization, a diffusion map 2D embedding was generated using R package density (Angerer et al., 2016). Next, a trajectory from Scara5+ fibroblast to myofibroblast was created using the function *addTrajectory* and visualized using the function *plotTrajectory*.

To identify TFs that drive this process, we first performed peak calling based on all fibroblasts using MACS2 (Zhang et al., 2008) to obtain specific peaks and then estimated motif deviation per cell using chromVAR (Schep et al., 2017). The deviation scores were normalized to allow for comparison between TFs. Next, we selected the TFs with high variance of deviation and gene activity score along the trajectory and calculated the correlation of TF activity and gene accessibility. This was done by using the function *correlateTrajectories* from ArchR.

### Experimental Validation of Runx1 by Immunofluorescence Staining

To validate the role of Runx1 in myofibroblast differentiation, we used immunofluorescence staining. Mouse kidney tissues were fixed in 4% formalin for 2 hours at RT and frozen in OCT after dehydration in 30% sucrose overnight. Using 5-10  $\mu\text{m}$  cryosections, slides were blocked in 5% donkey serum followed by 1-hour incubation of the primary antibody, washing 3 times for 5 minutes in PBS, and subsequent incubation of the secondary antibodies for 45 minutes. Following DAPI (4,6 – diamidino-2-phenylindole) staining (Roche, 1:10.000) the slides were mounted with ProLong Gold (Invitrogen, #P10144). Cells were fixed with 3% paraformaldehyde followed by permeabilization with 0,3% TritonX. Cells were incubated with primary antibodies and secondary antibodies diluted in 2% bovine serum albumin in PBS for 60 or 30 minutes, respectively. The following antibodies were used: anti-Runx1 (HPA004176, 1:100, Sigma-Aldrich), AF647 donkey anti-rabbit (1:200, Jackson Immuno Research). Images were acquired using a Nikon A1R con-focal microscope using 40X and 60X objectives (Nikon). Raw imaging data were processed using Nikon Software or ImageJ. Systematic random sampling was applied to the sub-sample of at least 3 representative areas per image of PDGFRbeGFP mice (n=3 mice per condition). Using QuPath nuclei were segmented and fluorescent intensity per nuclear size was measured of PDGFRbeGFP positive nuclei. The staining was performed by Christoph Kuppe.

### Computational Prediction of Runx1 Target Genes

After identifying and validating Runx1 as an important regulator, we next sought to computationally predict the target genes of Runx1. For this, we first performed co-accessibility analysis to link peak to genes as described in (Section 2.3.2). Next, we obtained the transcription start site (TSS) for each gene from reference genome mm10 and extended it by 250k bps for both directions. Then, we overlapped the peaks from fibroblasts and the TSS regions using function *findOverlaps* to identify putative peak-to-gene links. We next created 100 pseudo-bulk ATAC-seq profiles by assigning each cell to an interval along the trajectory of myofibroblast differentiation. The gene score matrix and peak matrix were aggregated according to the assignment to generate two pseudo-bulk data matrices. For each putative peak-to-gene link, we calculated the correlation between peak accessibility and gene activity. The p-values are computed using *t* distribution and corrected by the Benjamini-Hochberg method. For comparison, we also performed matrix imputation using the four top methods, i.e., scOpen, SCALE, MAGIC, and cisTopic, as evaluated by peaks recovering and computed the correlation based on the imputed matrix.

With each peak being associated with genes, we next sought to link Runx1 to its target genes. For this, we first performed footprinting with HINT-ATAC using the peaks obtained from above and pseudo-bulk ATAC-seq profiles. Next, we identified Runx1 binding sites using a motif matching approach. We defined the genes that have at least one footprint-support binding site of Runx1 in their associated peaks as Runx1 target genes. We then used the peak-to-gene correlation as a prediction between Runx1 and the target genes. This procedure was performed using the links estimated by different input data as described above, thus generating various predictions. To evaluate the results, we used the DE genes obtained from RNA-seq of Runx1 over-expression as true labels (see below),

### 4.3. Discussion

and computed the AUPR.

#### Experimental Validation of Runx1 Target Genes

To validate the target genes of Runx1, we first generated a human PDGFRb+ cell line that was isolated from the healthy part of the kidney cortex after nephrectomy as previously described in Kuppe et al. (2021) and over-expressed Runx1. Next, we extracted the RNA according to the manufacturer's instructions using the RNeasy Kit (QIAGEN) and performed sequencing on a NextSeq500/550 platform (Illumina) according to the manufacturer's protocols (Illumina, CA, USA).

For RNA-seq data analysis, we used the pipeline nf-core/rnaseq Patel et al. (2020). Briefly, reads were aligned to the hg38 reference genome using STAR Dobin et al. (2013) and gene expression was quantified with Salmon Patro et al. (2017). Differentially expressed genes were identified using DESeq2 Love et al. (2014). We used an adjusted p-value of 1e-05 and log2 fold change of 1 as thresholds to select the significant DE genes, which were used as true labels to evaluate the Runx1 target gene prediction (see above). GO enrichment analysis was performed with R package gprofiler2 and we showed results for biological process and pathways from Human Phenotype Ontology.

Nazanin Kabgani performed the cell culture experiments and generated the PDGFRb+ cell line together with Susanne Ziegler. Susanne Ziegler also performed the cloning for the over-expression. The RNA isolation and bulk RNA Seq library preparation were performed by Christoph Kuppe.

### 4.3 Discussion

---

In this chapter, we described the experimental framework to evaluate our computational method from technical and biological points of view. For technical validation, we first introduced a novel simulation algorithm for scATAC-seq and the described four real-world scATAC-seq datasets. The simulation data was used to test the hyper-parameters in scOpen and the real-world datasets were used benchmarking the performance. Next, we provided the full details of the execution and parameterization of the competing methods, i.e., eight imputation methods (MAGIC, SAVER, scImpute, DCA, cisTopic-impute, scBFA, SCALE, and imputePCA), three dimensionality reduction methods (cisTopic, SnapATAC, and Cusanovich2018). To achieve a comprehensive evaluation about data imputation and dimensionality reduction, we also introduced different metrics (Figure 4.1; Figure 4.3). For example, we compared the memory and running time requirements for imputation methods. Moreover, we proposed a novel approach to directly evaluate the imputation accuracy by estimating AUPR values for each cell against the true labels for the peaks. We also evaluated the distance accuracy between cells by using the cell labels. Furthermore, to fairly benchmark the clustering performance of the imputation methods, we employed two different clustering methods (i.e., k-medoids and hierarchical clustering), and performed the clustering analysis based on different inputs (i.e., PCA and t-SNE) and the number of clusters ( $k$  and  $k + 1$ ). Finally, we also tested whether scOpen can improve the downstream analysis of scATAC-seq. For this, we selected three different methods (i.e., Cicero, chromVAR, and scABC) and compared the distance and clustering accuracy between using raw count and scOpen imputed matrix as input for these methods (Figure 4.4).



For biological validation, we generated novel single-cell open chromatin data from the whole mouse kidney at different time points after unilateral ureteral obstruction (day 0, day 2, and day 10) using the droplet-based scATAC-seq protocol. We then applied scOpen on this dataset to produce a dimension-reduced matrix. Moreover, we also executed the competing method and compared the performance by using independent snRNA-seq data from the same mouse model as reference. Next, to investigate the cell-specific regulatory dynamic during fibrosis of kidney, we extended our method HINT-ATAC to infer differential TF activity between different cell types and time points. Finally, we performed additional analyses to gain novel biological insights about gene regulation of myofibroblast differentiation.



## Results

In the previous chapter, we introduced the experimental frameworks for evaluation of our computational method. In this chapter, we present the results generated by the analysis. The main structure of this chapter is the same as Chapter 4 for convenience. Specifically, we report the results for technical validation in Section 5.1 and the results for biological validation in Section 5.2.

### 5.1 Technical Validation

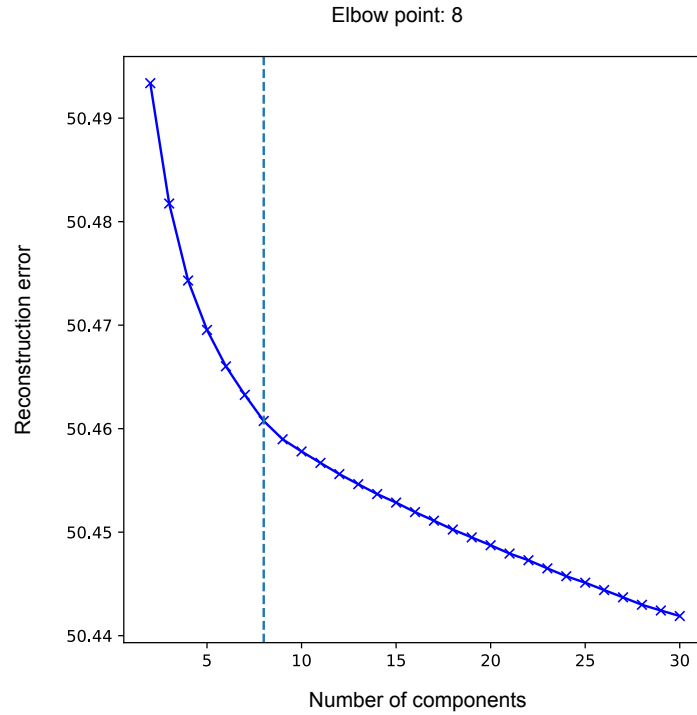
We here present the results of experiments for technically validating scOpen. First, we describe the outcomes of scOpen hyper-parameter selection as evaluated by using the simulated scATAC-seq data (Section 5.1.1). Next, we give the results of benchmarking scOpen and its competitors in terms of scATAC-seq data imputation (Section 5.1.2). Then, we describe the comparison results for dimensionality reduction methods in Section 5.1.3. Finally, we give the details of downstream analysis evaluation results (Section 5.1.4).

#### 5.1.1 scOpen Parameter Selection

We here used the simulation data generated as described in Section 4.1.1 to evaluate the parameters for our proposed computational approach, i.e., the number of components (or the matrix rank)  $k$  and the regularization parameter  $\lambda$  (Equation 3.10). First, we verified that our computational strategy for selecting the number of components, and we set  $\lambda = 1$  in this evaluation. The execution of scOpen on the simulation data automatically identified the optimal number of components as 8 (Figure 5.1). For comparison, we also applied scOpen with a range of  $k$  from 2 to 30. We then measured the imputation accuracy by AUPR and clustering accuracy by ARI as previously described in Section 4.1.4. Notably, we observed that the scOpen automatic procedure for components selection obtained the best results in terms of imputation accuracy (Figure 5.2a-b; Appendix Table A.2). Moreover, we observed a similar clustering accuracy between the estimated number of components and the best results (Figure 5.2c-d; Appendix Table A.3). Together, these results indicated our computational approach is able to produce a good estimation for the number of components and we will apply this strategy in the following benchmarking analysis.

Next, we evaluated the regularization parameter  $\lambda$ . For this, we applied scOpen with a number of values of  $\lambda$  from 0 to 4. We benchmarked the performance by imputation accuracy as measured by AUPR and clustering accuracy as measured by ARI, respectively. The results revealed that a value of 1 is optimal in the imputation problem (Figure 5.3a; Appendix Table A.4), while values in the range

## 5.1. Technical Validation



**Figure 5.1: Estimation of the number of components for the simulation data.** The estimated number of components for the simulation data using elbow detection approach. The x-axis represents the number of components, and the y-axis represents the reconstruction error. The dashed line indicates the detected elbow point.

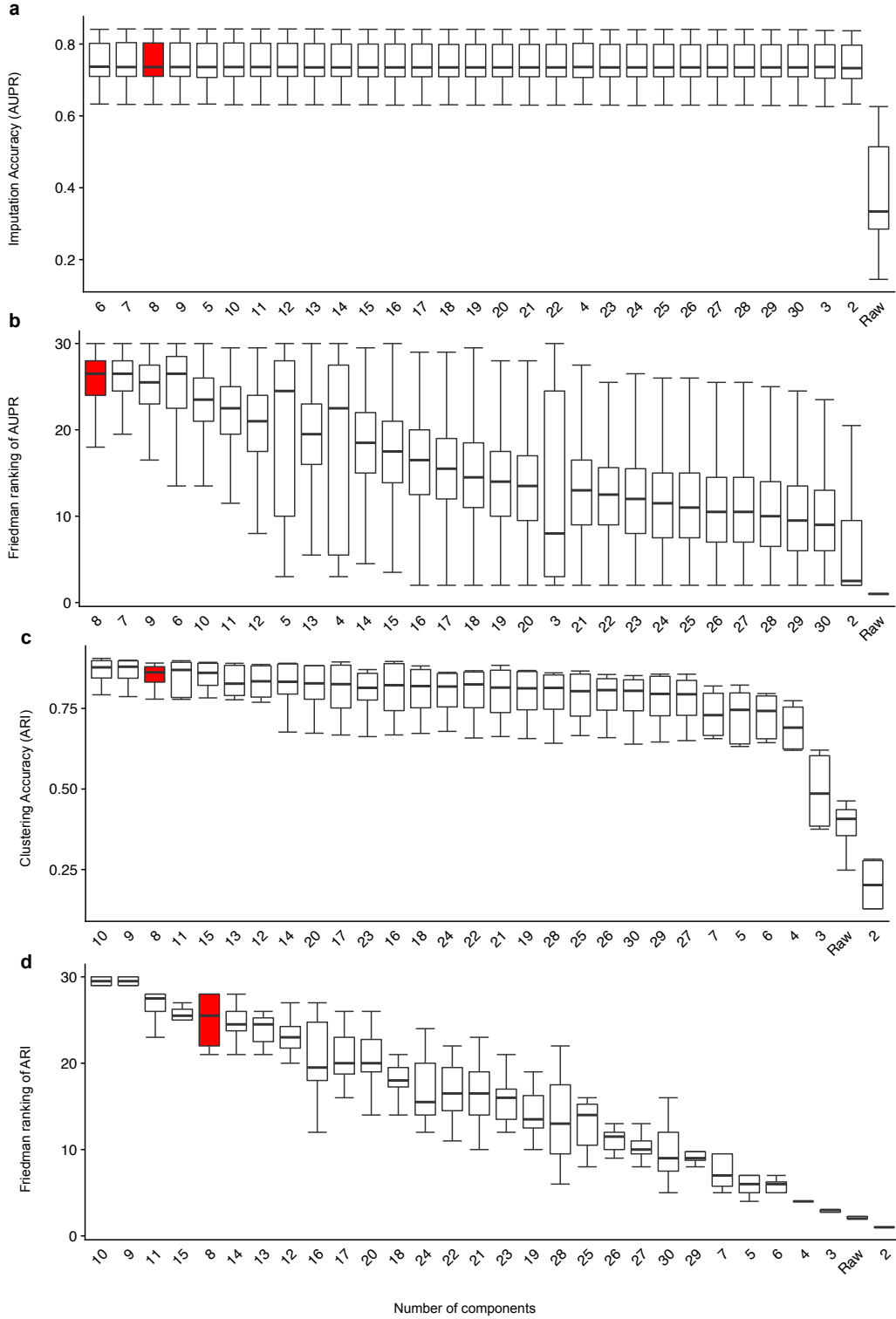
[0, 1] are optimal for the clustering problem (Figure 5.3b; Appendix Table A.5). Collectively, these results highlighted the importance of applying regularization for scATAC-seq data imputation. In the following benchmarking, we used the  $\lambda = 1$  as default for scOpen.

### 5.1.2 Evaluation of Imputation Methods

Having determined the optimal hyper-parameters used by scOpen, we here performed a comprehensive evaluation to compare the performance of scOpen and its competitors in terms of scATAC-seq data imputation. We executed the selected imputation methods as described in Section 4.1.3 on the benchmarking data (Section 4.1.1) and evaluated the results from different aspects as detailed below.

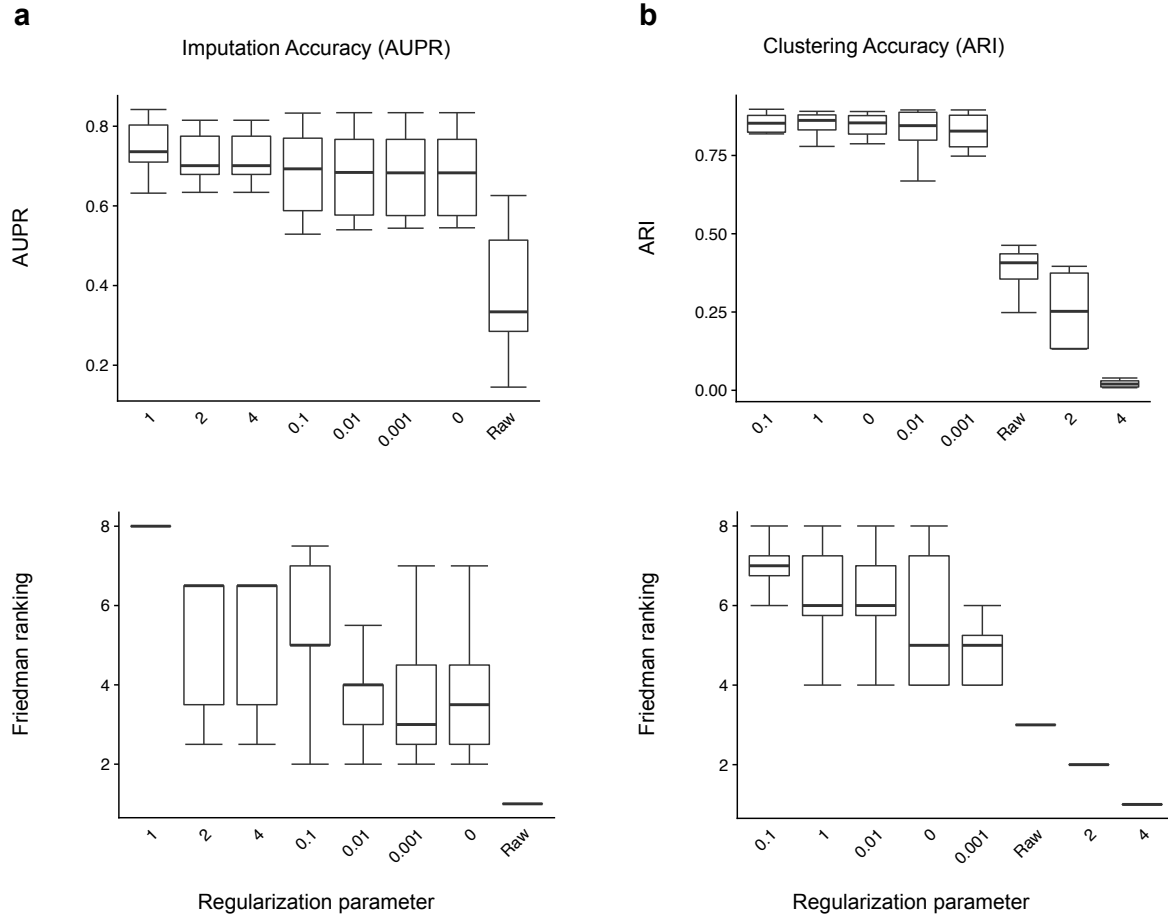
#### Benchmarking Requirements of Memory and Time

We first evaluated the memory and running time requirements of imputation methods (Figure 4.1.4). Overall, we observed that scOpen had the lowest memory requirements, i.e., it required at least 2 fold less memory as compared to cisTopic, MAGIC and SCALE (Figure 5.4a). Moreover, it had a maximum requirement of 16GB on the PBMC dataset, meaning that it is possible to run scOpen on a commonly available laptop even for a large dataset with about 10,000 cells (see Table 4.1). Regarding computing time, MAGIC was the fastest followed by SCALE and scOpen. These were the only methods performing imputation for the large *PBMC* dataset (10k cells vs. 100k peaks) in less than 3 hours (Figure 5.4b), while imputePCA, SAVER and DCA failed to execute at the *PBMC*



**Figure 5.2: Evaluation of the number of components using the simulation data.** **a**, Boxplot comparing the imputation accuracy of scOpen by using different number of components. The x-axis represents the number of components, and the y-axis represents the AUPR. Methods are ranked by the average AUPR. The estimated components number is highlight with red color. **b**, Same as **a** for the Friedman ranking of AUPR. **c**, Same as **a** for clustering accuracy as measured by ARI. **d**, Same as **a** for the Friedman ranking of ARI.

## 5.1. Technical Validation

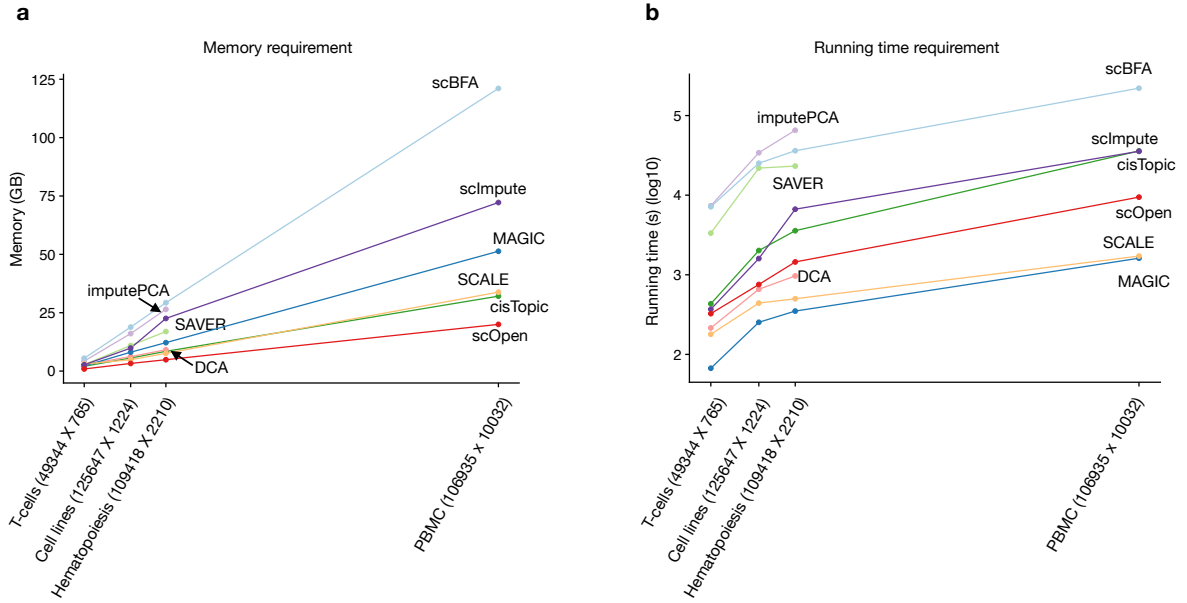


**Figure 5.3: Evaluation of the regularization parameter using the simulation data.** **a**, Box plots comparing the imputation accuracy of scOpen by using different regularization parameters as measured by AUPR (upper) and the Friedman ranking (lower). Methods are ranked by the average AUPR or ranking score. **b**, Same as **a** for clustering accuracy.

dataset. These results indicated that scOpen had a good scalability.

### Benchmarking Imputation Accuracy

We next tested if imputation methods can improve the recovery of true open chromatin (OC) regions. We termed this metric as imputation accuracy. For this, we created the true positive and negative OC labels for each single-cell in the benchmarking datasets as described in Section 4.1.4. Next, we computed the AUPR and performed the non-parametric Friedman test with the Nemenyi post-hoc test. Notably, we observed that scOpen significantly outperformed most of the competing methods by presenting the highest average Friedman ranking of the AUPR across all benchmarking datasets (Figure 5.5; Appendix Table A.6 - Appendix Table A.9). The combined ranking indicated that SCALE and MAGIC were the runner-up methods (Appendix Figure A.1a). Moreover, we also evaluated the influence of the number of cells per cluster on the AUPR. Despite an overall decrease in AUPR with sample size, we observed that top performing methods (i.e., scOpen, SCALE, and MAGIC) were less sensible to the cell numbers (Figure 5.6).



**Figure 5.4: Evaluation of the memory and running time requirements for imputation methods.**

**a**, The memory requirements (y-axis) of the imputation/denoising methods on benchmarking datasets. The x-axis represents the number of elements of the input matrix (number of OC regions by cells). **b**, Same as **a** for running time requirements.

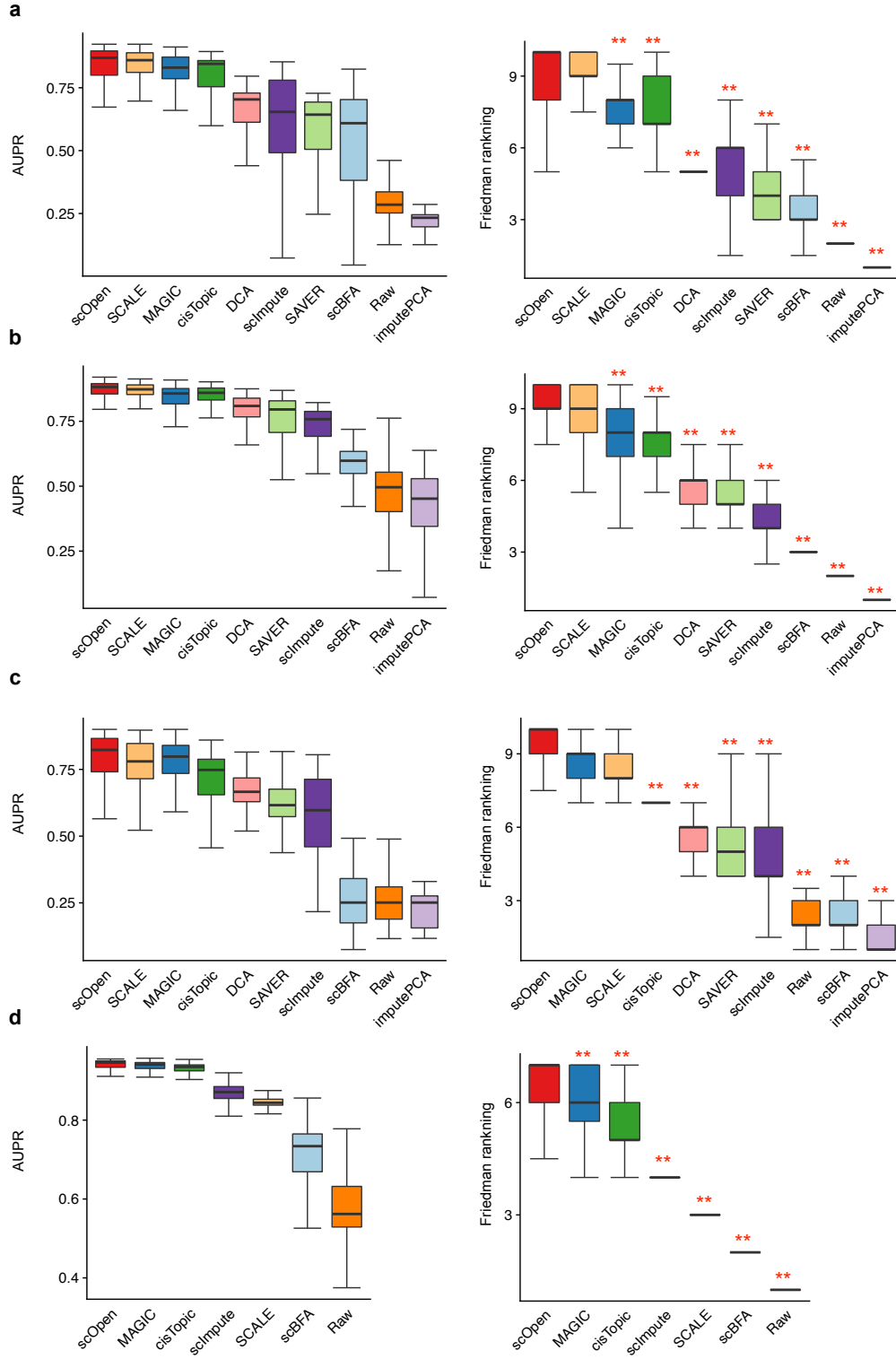
### Benchmarking Distance Accuracy

We then investigated the impact of imputation on the estimation of distances between cells. We estimated the cell-to-cell distance as  $1 - \text{Pearson correlation}$  for each dataset using the imputed matrix. We evaluated the results using silhouette score regarding the agreement with known cell labels and performed the non-parametric Friedman test with the Nemenyi post-hoc test to find out if there are any significant differences as previously described (Section 4.1.4). We observed that scOpen significantly outperformed all competing methods in all benchmarking datasets by presenting the highest average Friedman ranking of the silhouette score (Figure 5.7; Appendix Table A.10 - Appendix Table A.13). The combined results indicated that cisTopic and MAGIC were the runner-up methods (Appendix Figure A.1b).

### Benchmarking Clustering Accuracy

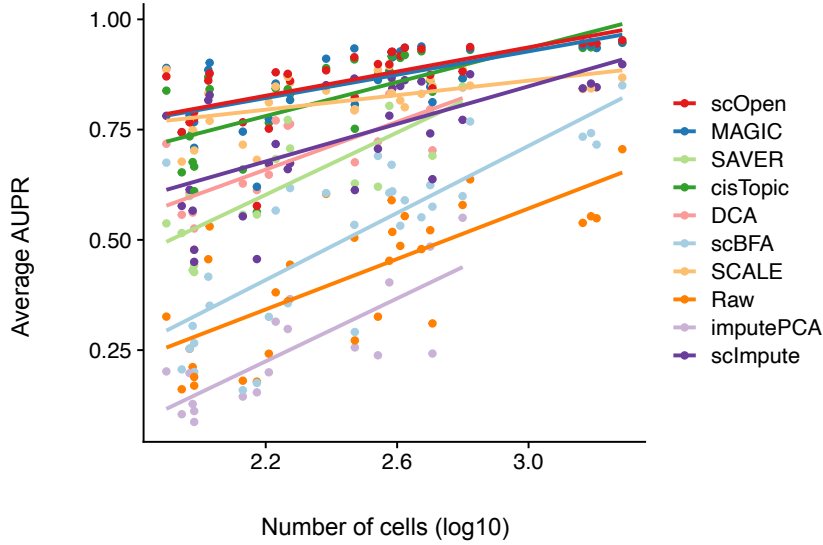
Finally, we evaluated the imputation results based on clustering. For each dataset, we clustered the cells using the imputed matrix according to Section 4.1.4 and benchmarked the clustering performance by calculating ARI. We observed that scOpen was the best in the *Hematopoiesis*, *T cells*, and multi-omics *PBMC* datasets and the second best for the *Cell lines* dataset (Figure 5.8; Appendix Table A.14 - Appendix Table A.17). When considering the combined ranking, scOpen was the best performer followed by MAGIC and cisTopic (Appendix Figure A.1c). The discriminative power of scOpen was also supported by UMAP Becht et al. (2019) projections of these datasets, which provides a clear separation of the majority of cell labels for each dataset (Figure 5.9 and Figure 5.10). Altogether, these results supported that scOpen outperformed the state-of-the-art imputation methods

## 5.1. Technical Validation



**Figure 5.5: Evaluation of the imputation methods based on imputation accuracy.** **a**, Box-plot comparing the imputation accuracy of the imputation methods as measured by AUPR (left) and the Friedman ranking of AUPR (right) for *Cell line* dataset. The asterisk and the two asterisks, respectively, mean that the method is outperformed by scOpen with significance levels of 0.05 and 0.01. **b**, Same as **a** for *Hematopoiesis* dataset. **c**, Same as **a** for *T cells* dataset. **d**, Same as **a** for *PBMC* dataset.





**Figure 5.6: Association between the number of cells and AUPR.** Scatter plot associating the average AUPR and number of cells for each cell type in all benchmarking datasets. Each dot represents a cell type and color refers to method. The x-axis represents the number of cells for each cell type. The trend line was fitted for each method.

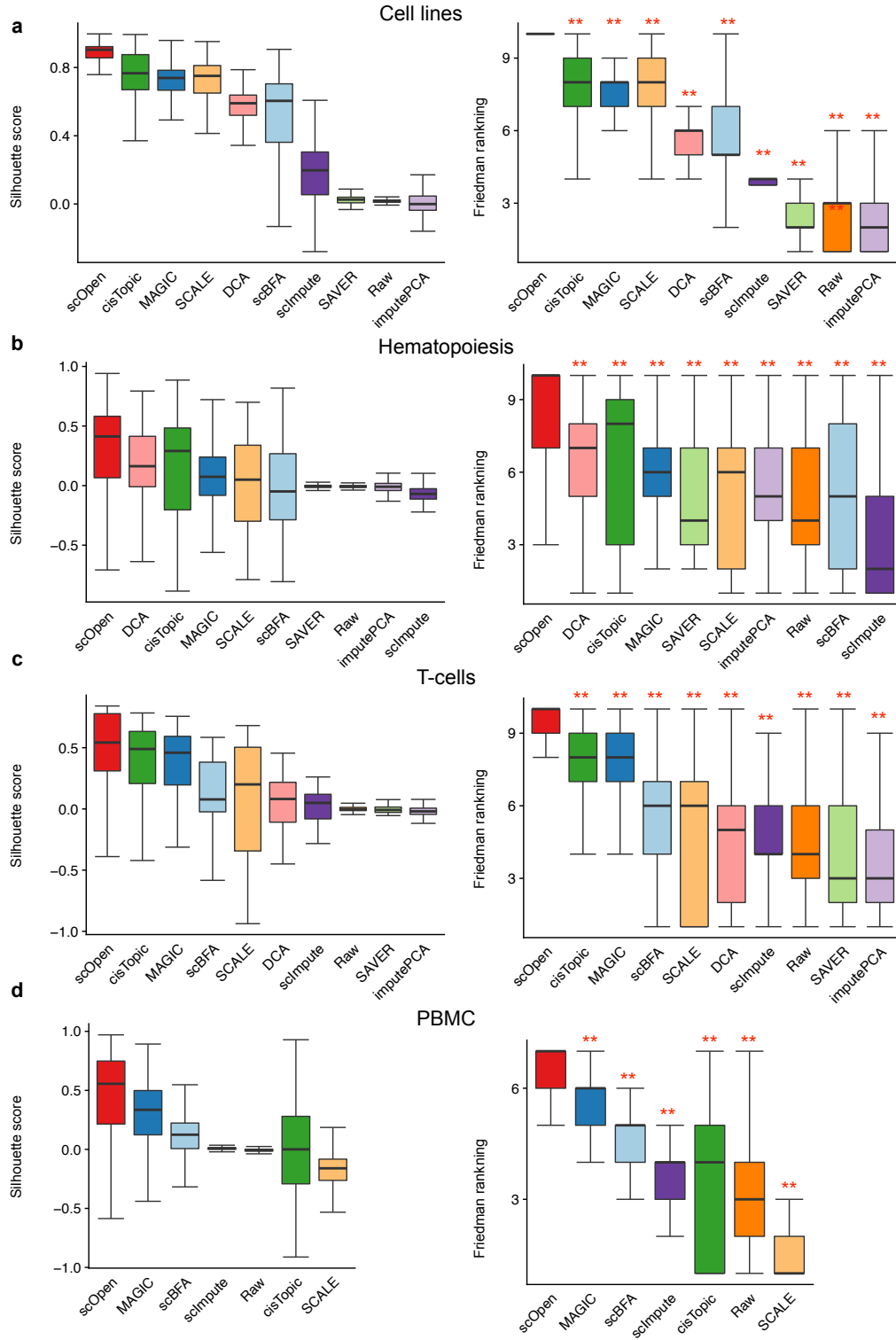
while providing the lowest memory footprint and an above average running time performance.

### 5.1.3 Evaluation of Dimensionality Reduction Methods

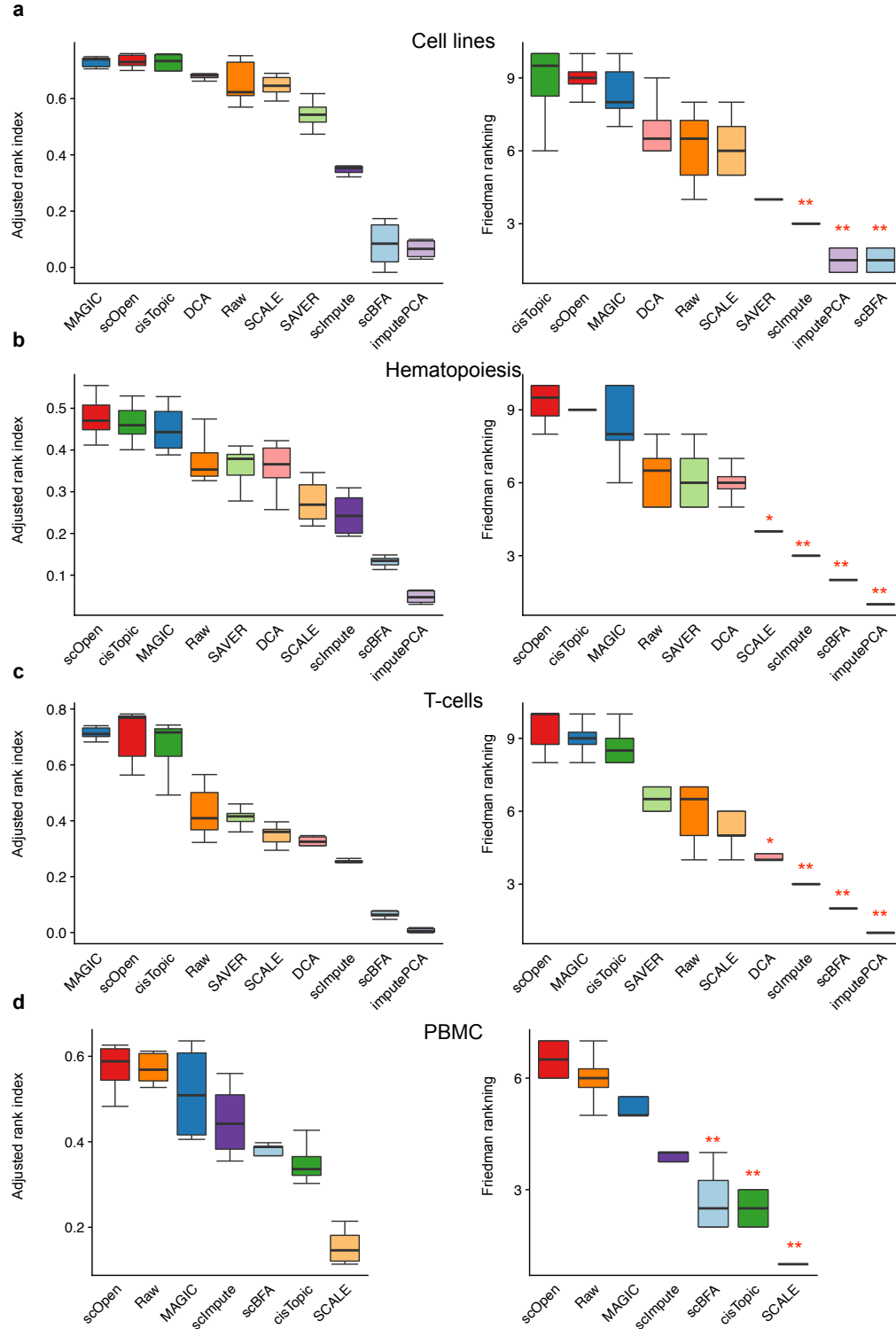
Another relevant question was to compare scOpen with top performing state-of-the-art scATAC-seq dimensionality reduction methods: cisTopic, SnapATAC and Cusanovich2018 (Section 2.4.3). Here, we executed the SnapATAC and Cusanovich2018 for each of the benchmarking datasets to produce a low-dimensional representation of the cells as described in Section 4.1.4. We also evaluated the use of both reduced and imputed matrices for scOpen and cisTopic, as these methods provided both types of representations. We first compared the distance accuracy using the silhouette score with the non-parametric Friedman-Nemenyi test. Interestingly, we observed that either scOpen imputed or low-dimension matrices significantly outperformed the competitors by presenting the highest score in 3 out of the benchmarking datasets (Figure 5.11; Appendix Table A.18 - Appendix Table A.21). Moreover, both scOpen matrix representations tied as first in the combined rank (Appendix Figure A.2). cisTopic, which was the runner-up method, performed well in datasets *Cell line*, *Hematopoiesis*, and *T-cells* but poorly for *PBMC*.

Next, we evaluated the clustering performance of these methods. It is worth pointing out that each competing method has an accompany method for clustering as proposed in the original paper, e.g., graph-based clustering for SnapATAC (Fang et al., 2021) and density-based clustering for cisTopic and Cusanovich2018 (González-Blas et al., 2019). Therefore, we used these clustering approaches instead of the k-medoids and hierarchical clustering methods to achieve a fair comparison (Section 4.1.4). We observed that scOpen performed best on *Cell line* and *Hematopoiesis* datasets and was ranked first/second in the combined rank (Figure 5.12). Overall, this analysis indicated that both reduced dimension and imputed scOpen matrices obtained the best overall results for distance and

## 5.1. Technical Validation

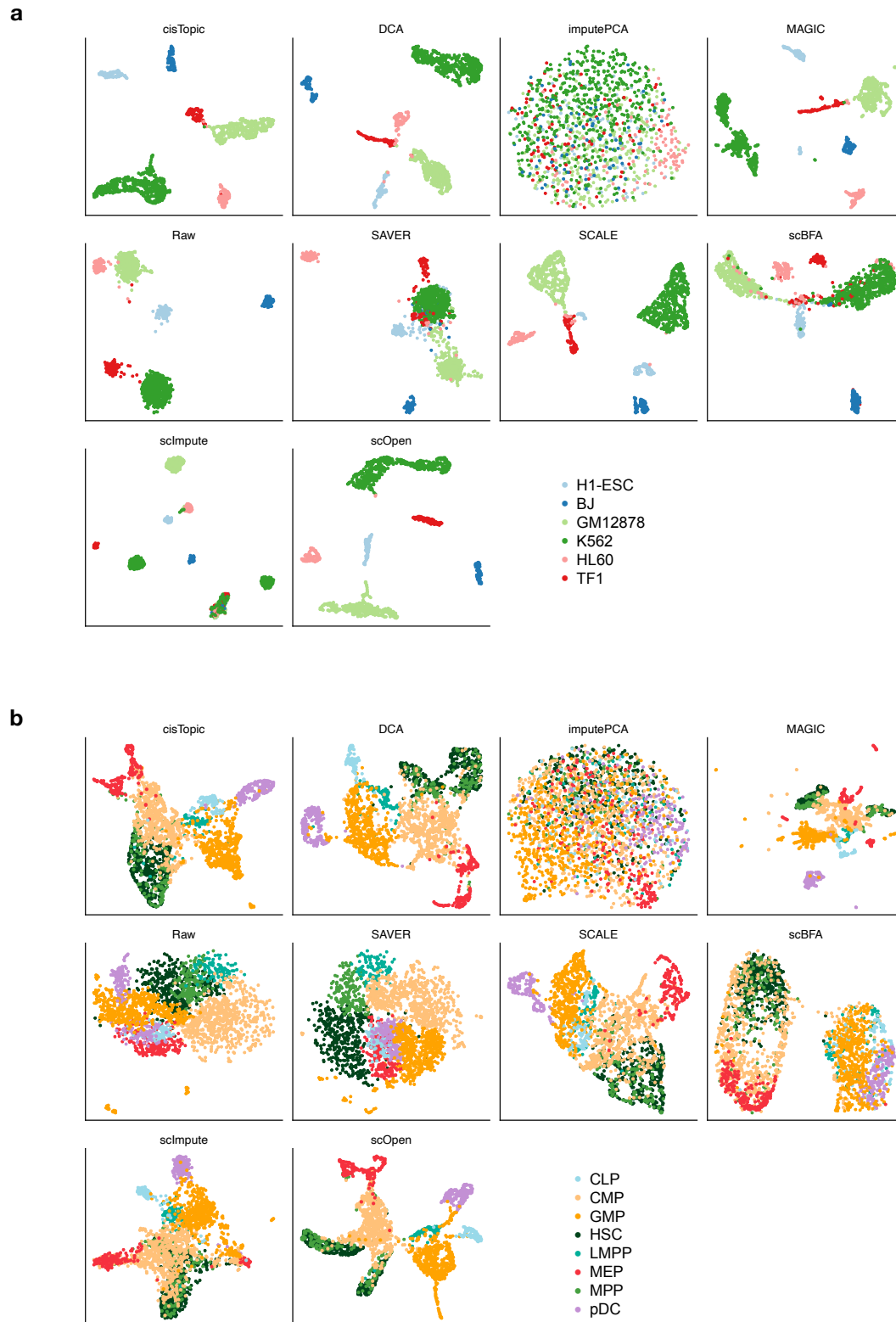


**Figure 5.7: Evaluation of the imputation methods based on distance accuracy.** **a**, Boxplot comparing the distance accuracy of the imputation methods as measured by silhouette score (left) and the Friedman ranking of silhouette score (right) for *Cell line* dataset. The asterisk and the two asterisks, respectively, mean that the method is outperformed by scOpen with significance levels of 0.05 and 0.01. **b**, Same as **a** for *Hematopoiesis* dataset. **c**, Same as **a** for *T cells* dataset. **d**, Same as **a** for *PBMC* dataset.

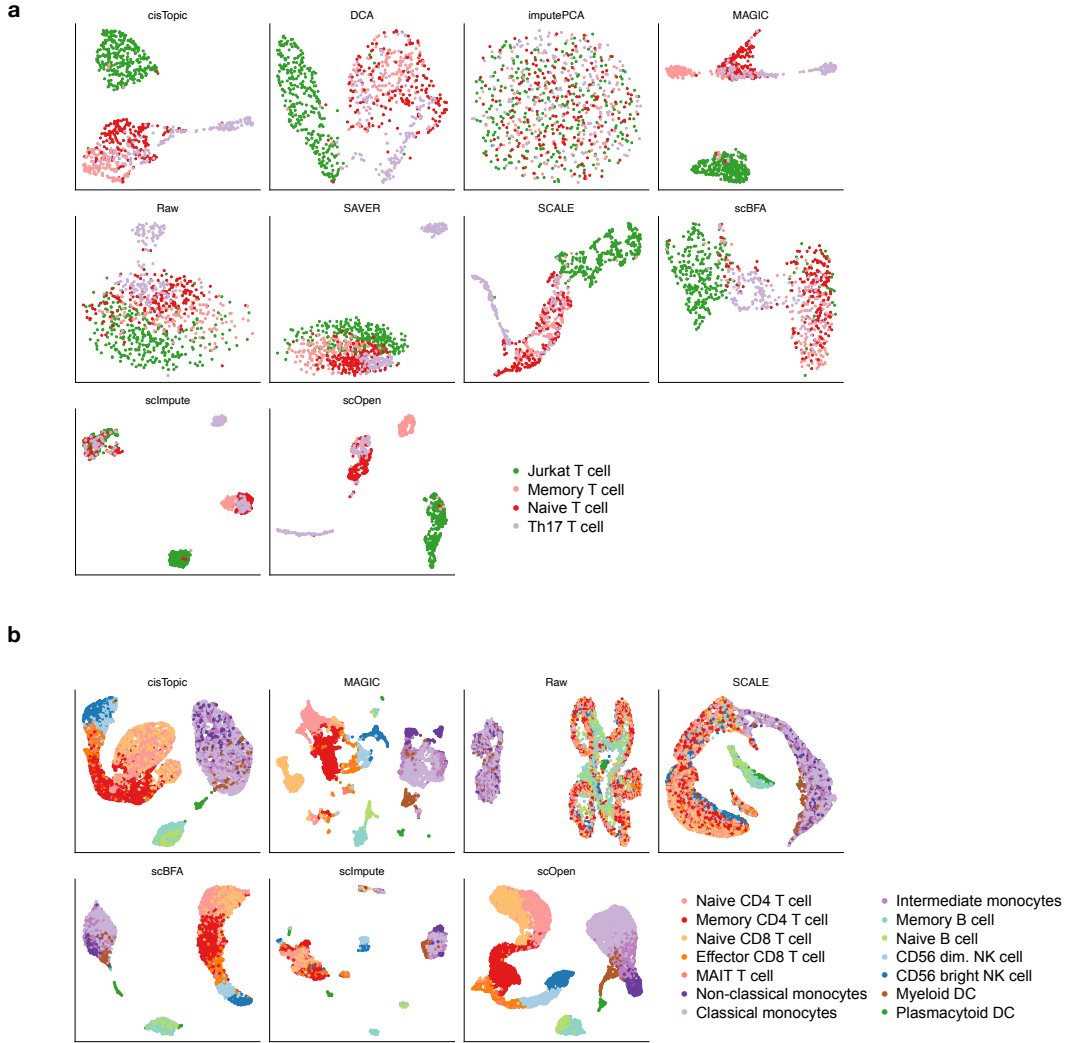


**Figure 5.8: Evaluation of the imputation methods based on clustering accuracy.** **a**, Boxplot comparing the clustering accuracy of the imputation methods as measured by rand adjust index (left) and the Friedman ranking of rank adjust index (right) for *Cell line* dataset. The asterisk and the two asterisks, respectively, mean that the method is outperformed by scOpen with significance levels of 0.05 and 0.01. **b**, Same as **a** for *Hematopoiesis* dataset. **c**, Same as **a** for *T cells* dataset. **d**, Same as **a** for *PBMC* dataset.

## 5.1. Technical Validation



**Figure 5.9: Visualization of imputation methods on benchmarking datasets.** **a**, UMAP embedding of scOpen, cisTopic, DCA, MAGIC, SAVER, scImpute, imputePCA and the raw data for *Cell line* dataset. **b**, Same as **a** for *Hematopoiesis* dataset.



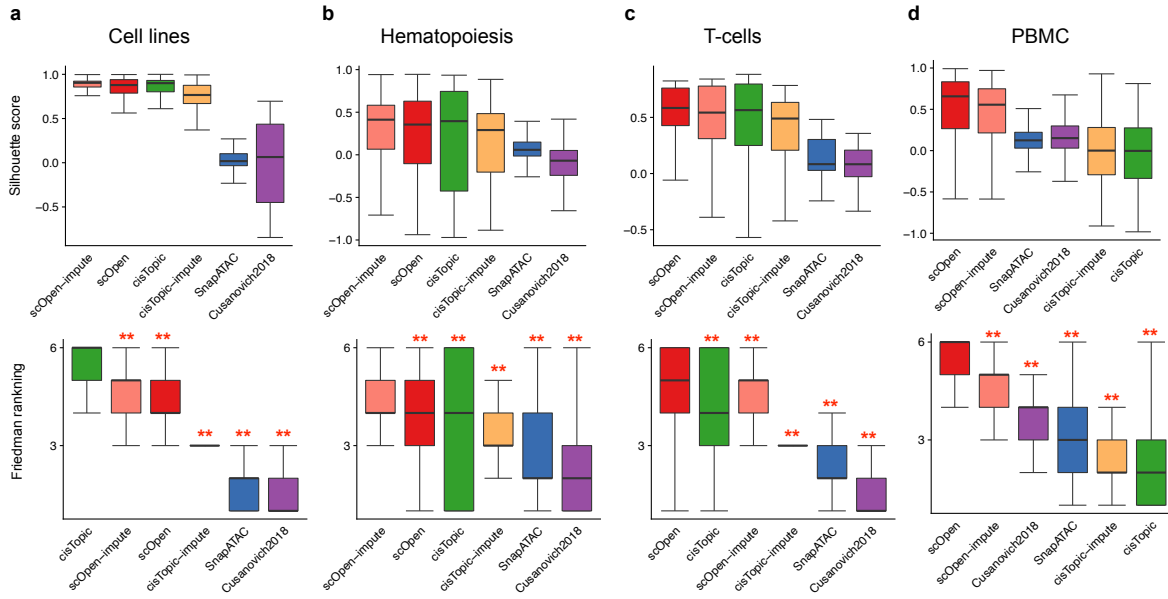
**Figure 5.10: Visualization of imputation methods on benchmarking datasets. a**, UMAP embedding of scOpen, cisTopic, DCA, MAGIC, SAVER, scImpute, imputePCA and the raw data for *T cells* dataset. **b**, Same as **a** for *PBMC* dataset.

clustering representations on benchmarking datasets. Of note, the low dimensional matrix reduced the memory footprint on the clustering by > 1000 fold in comparison to using full imputed matrices, serving as an alternative for clustering of the high-dimensional datasets.

#### 5.1.4 Evaluation of Downstream Analysis Methods

Next, we tested the benefit of using scOpen estimated matrices as input for scATAC-seq computational pipelines, which have as objective the identification of regulatory features associated with single cells (chromVAR (Schep et al., 2017)), estimation of gene activity scores and DNA-interactions (Cicero (Pliner et al., 2018)), or a clustering method tailored for scATAC-seq data (scABC (Zamanighomi et al., 2018)). Both chromVAR and Cicero first transformed the scATAC-seq matrix to either transcription factors and genes feature space, respectively. Clustering was then performed using the standard pipelines from each approach. We compared the distance and clustering accuracy of these methods using either raw or scOpen estimated matrices as input (Figure 4.4). In all combinations of

## 5.1. Technical Validation

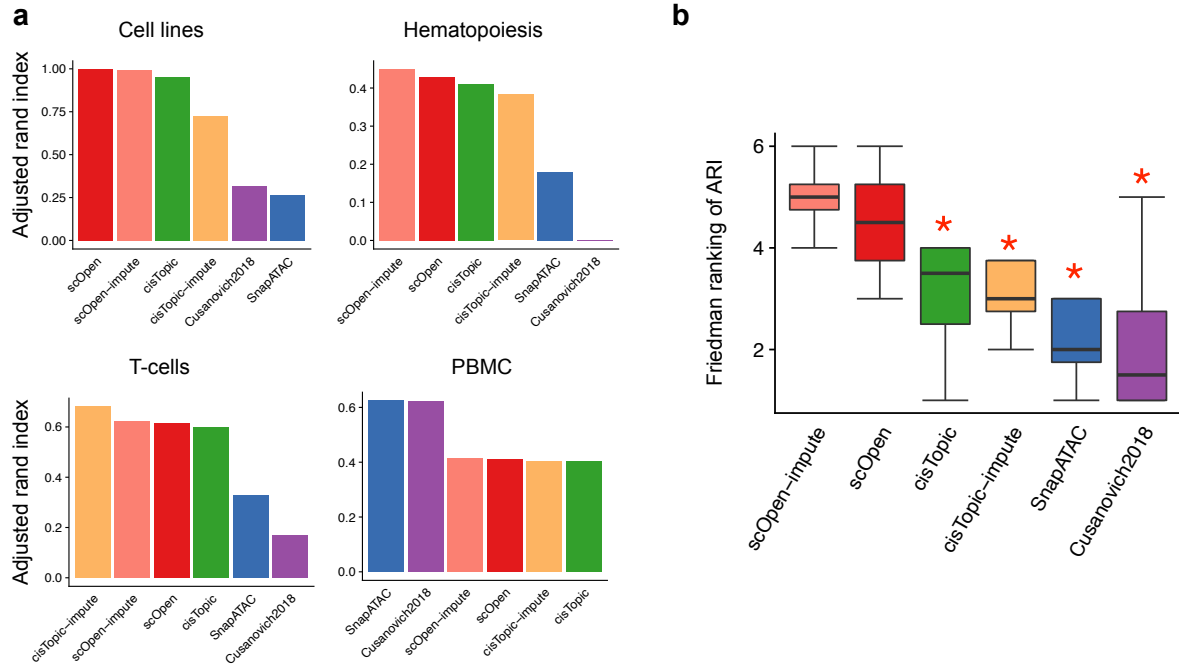


**Figure 5.11: Evaluation of the dimensionality reduction methods using distance accuracy.** Box-plot comparing the distance accuracy of the imputation methods for each benchmarking dataset as measured by silhouette score (upper) and Friedman ranking of silhouette score (lower).

methods and datasets, we observed a higher or equal silhouette and ARI whenever a scOpen matrix was provided as input (Figure 5.13). These results were also reflected in the UMAP visualization of the transformed matrices as generated by chromVAR and Cicero using scOpen imputed or raw count matrix as input (Figure 5.14). Altogether, these results indicated that the use of scOpen estimated matrices improved downstream analysis of state-of-the-art scATAC-seq methods.

### 5.1.5 Evaluation of Co-accessibility Analysis

In order to estimate the gene activity score from single-cell open chromatin data, Cicero first predicted co-accessible pairs of DNA regions in groups of cells, which potentially form *cis*-regulatory interactions as demonstrated in Figure 4.5. Here, we also evaluated the predicted interactions made by Cicero. For this, we applied the imputation methods to obtain the imputed matrices for human lymphoblastoid cells (GM12878) and then ran Cicero to generate the prediction by using either imputed or raw count matrices as described in Section 4.1.4. Next, we compared the prediction with Hi-C and ChIA-PET data from this cell type as provided by Pliner et al. (2018). Both protocols quantified the number interactions between genome loci, thus providing the true labels for evaluation. We calculated the AUPR values and odds ratios using these true labels. Notably, we observed that the imputation improved the detection of interactions globally and scOpen achieved the best results as measured by AUPR and odds ratios (Figure 5.15; Appendix Figure A.3). To evaluate the impact on the number of cell on these predictions, we have down-sampled the data to only consider 50% or 25% of cells. We observed a residual decrease in the AUPR of scOpen for 25% of cells (Appendix Figure A.5). This supported that chromatin conformation prediction works well even for cell types



**Figure 5.12: Evaluation of the dimensionality reduction methods using clustering accuracy. a,** Barplot comparing the clustering accuracy of the imputation methods for each benchmarking dataset as measured by ARI. **b,** the Friedman ranking of dimensionality reduction methods in terms of the average adjusted rand index for each benchmarking dataset. Methods are ordered by median value of ranks. Wilcoxon Rank Sum test was used to compare scOpen-impute with other methods. The asterisk means that the method is outperformed by scOpen with significance level of 0.05.

with low abundance.

The power of scOpen imputation was clear when checking the individual locus (Figure 5.16), as previously presented by Cicero Pliner et al. (2018). This is evident when contrasting accessibility scores between pairs of peak-to-peak links supported by Hi-C predictions (Figure 5.17; Appendix Figure A.4). scOpen obtained highly correlated accessibility scores, while other imputation methods showed quite diverse association patterns.

## 5.2 Biological Validation

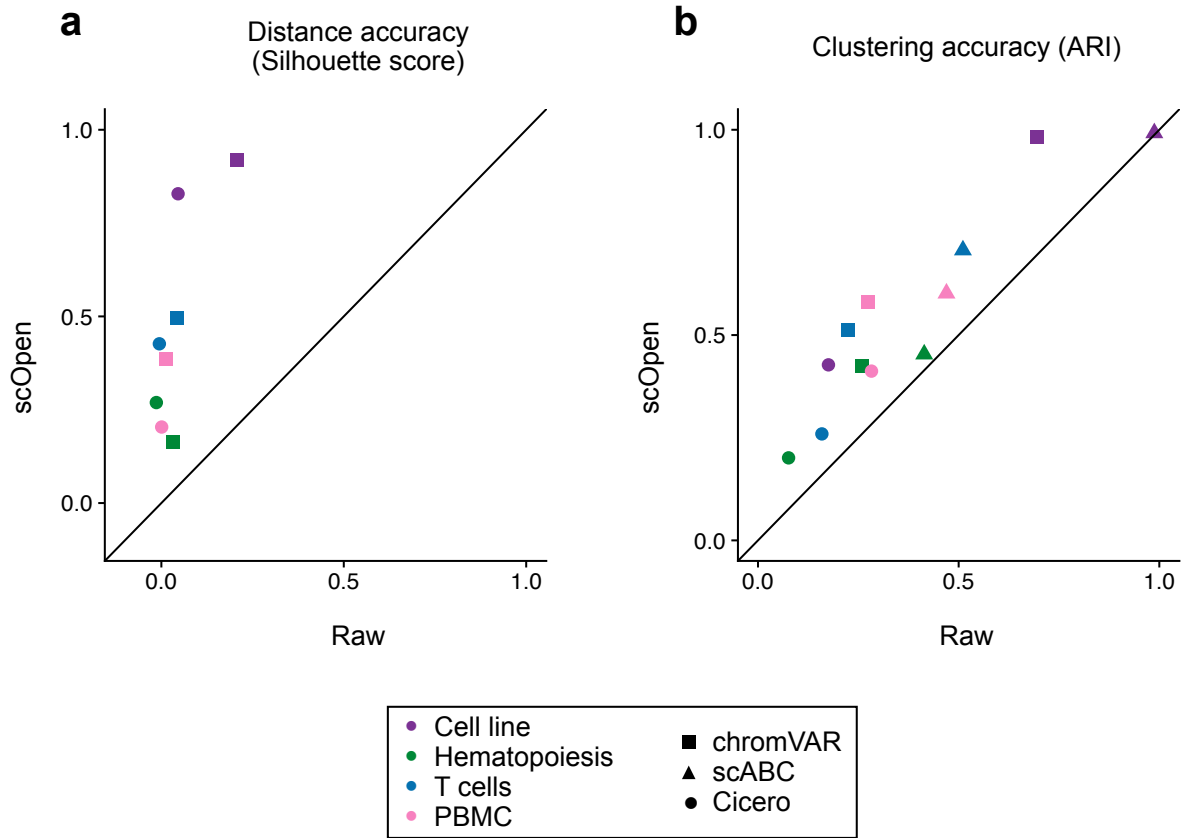
In this section, we provide the results obtained by performing the experiments described in Section 4.2.

### 5.2.1 Applying scOpen to Complex Disease scATAC-seq Data

#### Computational Analysis of scATAC-seq Data with scOpen

We first evaluated scOpen in its power to improve the detection of cells in a complex disease dataset. For this, we generated scATAC-seq for whole mouse kidney from C57Bl6/WT mice in homeosta-

## 5.2. Biological Validation



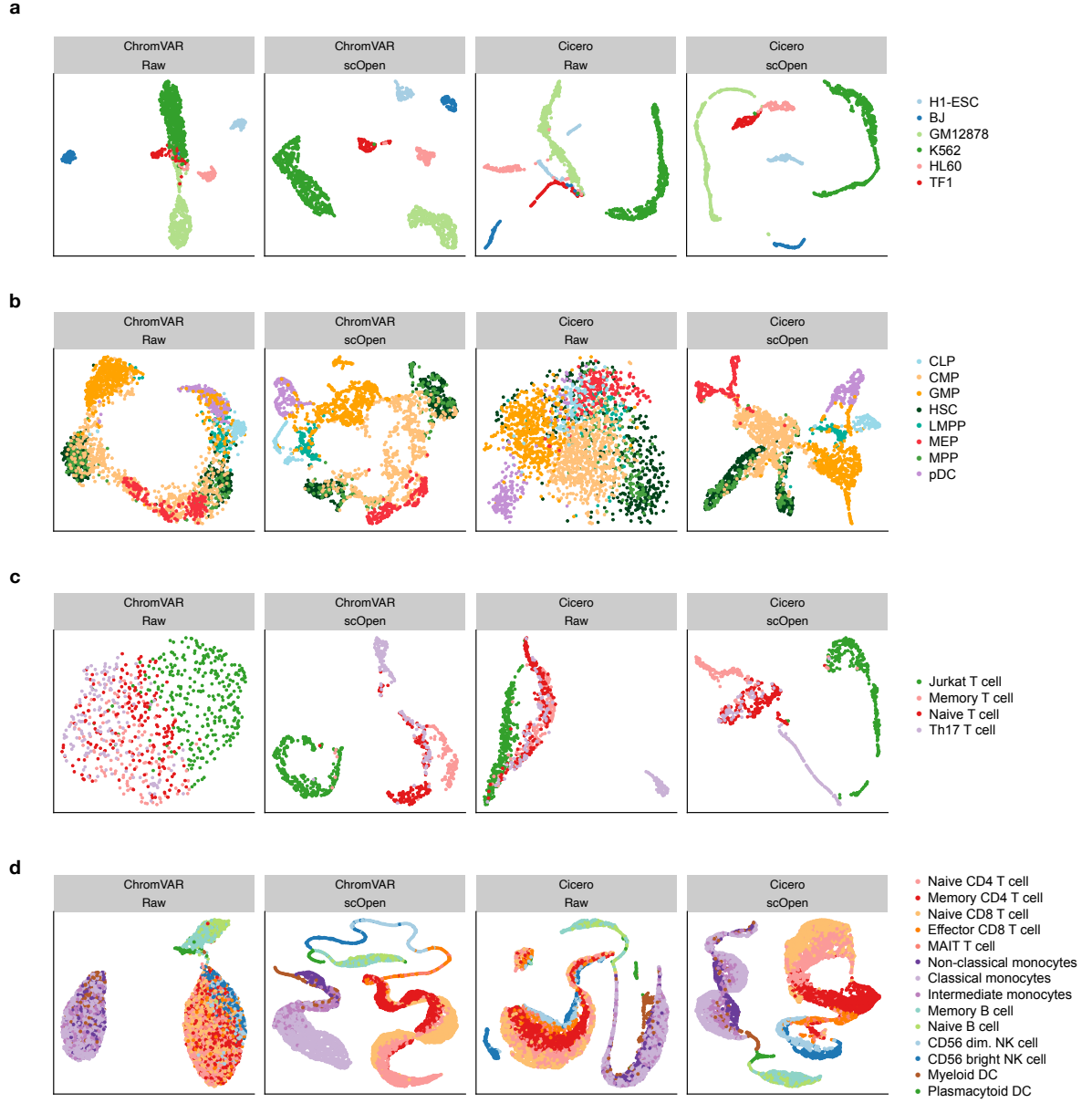
**Figure 5.13: Evaluation of the downstream analysis methods.** **a**, Scatter plot comparing silhouette score of datasets by providing raw (x-axis) and scOpen estimated matrices (y-axis) as input for Cicero and chromVAR. Colors represent datasets and shapes represent methods. scABC is not evaluated as it does not provide a space transformation. **b** Same as **a** for clustering results (ARI) of Cicero, chromVAR and scABC.

sis (day 0) and at two time points after injury with fibrosis: 2 days and 10 days after unilateral ureteral obstruction (UUO) as described in Section 4.2.1. After data preprocessing and quality controlling, we recovered a total of 30,129 cells with average of 13,933 fragments per cell, a fraction of reads in promoter of 0.46, and high reproducibility between biological duplicates (Appendix Figure A.6). We next performed peak calling and detected 150,593 peaks. These peaks were used as features to construct a raw scATAC-seq count matrix with high dimensionality and sparsity (4.2% of non-zeros).

Next, we performed data integration using Harmony algorithm (Korsunsky et al., 2019) to remove batch effects. For comparison, we also used dimension reduced matrices from either Cusanovich2018 (LSI), cisTopic, SnapATAC, or scOpen. We annotated the scATAC-seq profiles using single nuclei RNA-seq (snRNA-seq) data of the same kidney fibrosis model from an independent study (Wu et al., 2019) via label transfer (Stuart et al., 2019) to serve as true cell labels. We then evaluated the batch correction results based on clustering accuracy as measured by ARI and distance accuracy as measured by silhouette score (Section 4.1.4).

Notably, we observed that clusters based on scOpen were more similar to the transferred labels (higher ARI) than clusters based on competing methods (Figure 5.18a). Furthermore, scOpen

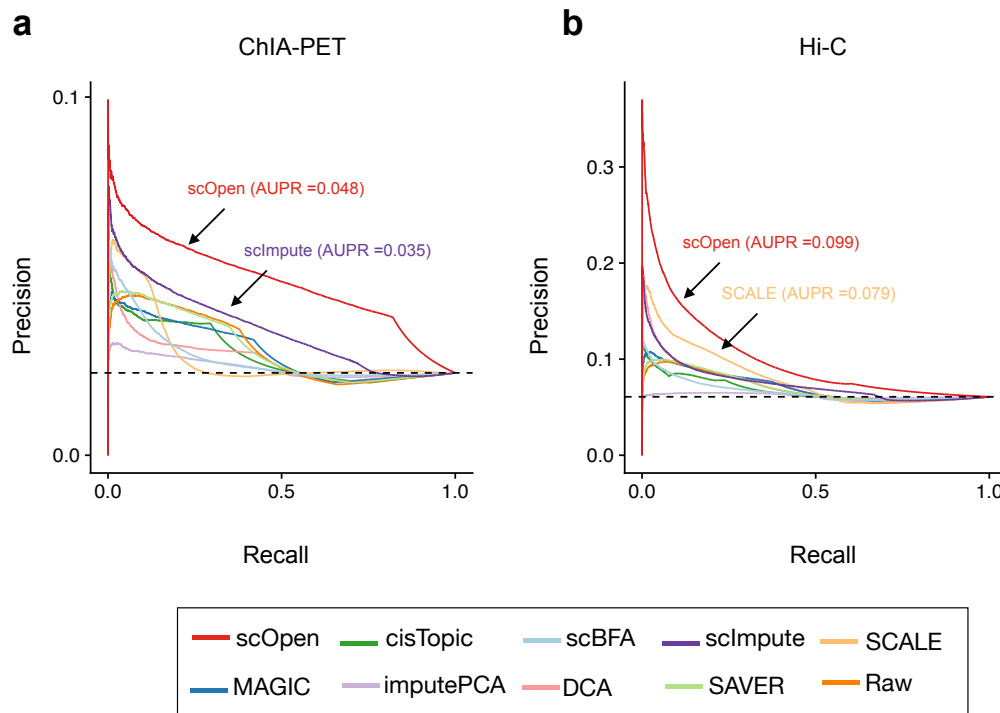




**Figure 5.14: Visualization of the downstream analysis methods.** **a**, UMAP embedding of chromVAR and Cicero transformed data using either raw or scOpen estimated matrix as input for *Cell line* dataset. **b**, Same as **a** for *Hematopoiesis* dataset. **c**, Same as **a** for *T cells* dataset. **d**, Same as **a** for *PBMC* dataset.

also provided better distance metrics and visualization than competing methods (Figure 5.18b; Appendix Figure A.7). These results supported the discriminative power of scOpen in this large and complex scATAC-seq dataset.

## 5.2. Biological Validation

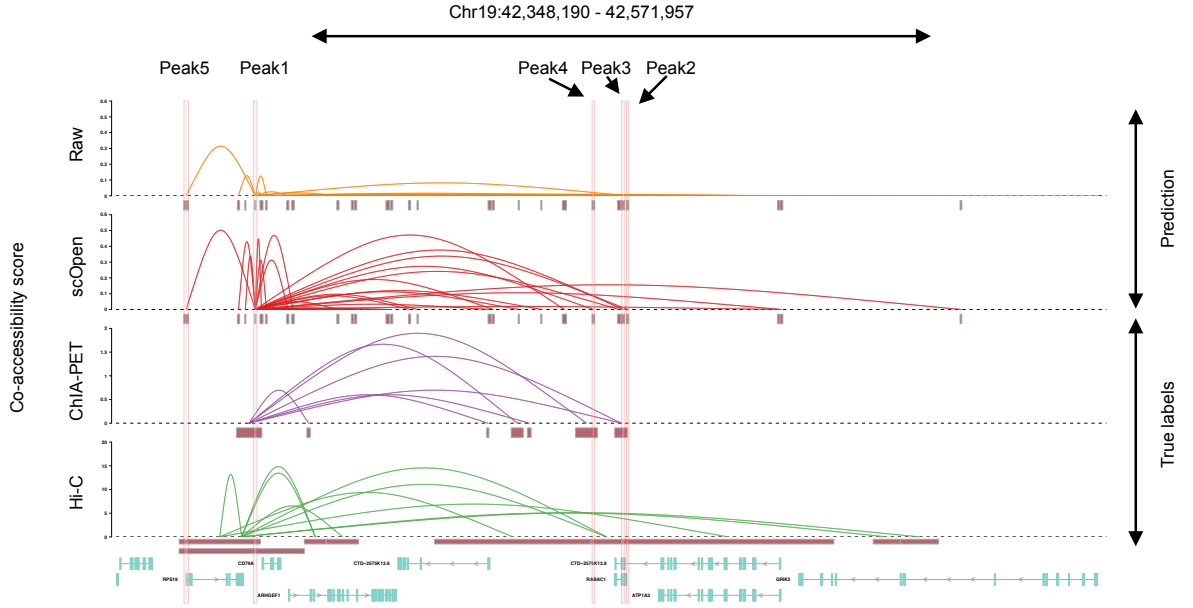


**Figure 5.15: Evaluation of Cicero predicted DNA regulatory elements interactions.** **a**, Precision-recall curves showing evaluation of the predicted links on GM12878 cells using raw and imputed matrix as input. We used data from pol-II ChIA-PET as true labels. Colors refer to methods. We reported the AUPR values for the top 2 methods. **b**, Same as **a** by using Hi-C data as true labels.

## Cell Annotation

Next, we annotated the clusters of scOpen by combining known marker genes and transferred labels after removing doublets with ArchR (Granja et al., 2021)<sup>1</sup>. Notably, we identified all major kidney cell types, including proximal tubular cells (PT), distal/connecting tubular cells (DCT), collecting duct and loop of Henle, endothelial cells (EC), fibroblasts as well as the rare populations of podocytes (Pod) and lymphocytes (Lymphoid) (Figure 5.19a; Appendix Figure A.8). Lymphocytes were not described in the previously scRNA-seq study (Wu et al., 2019), which supported the importance of annotation of scATAC-seq clusters independently of scRNA-seq label transfer. Of particular interests were cell types with population changes during progression of fibrosis (Figure 5.19b; Appendix Figure A.9). We observed an overall decrease of normal proximal tubular, glomerular, and endothelial cells and an increase of immune cells as expected in this fibrosis model with tubule injury, the influx of inflammatory cells, and capillary loss (Bábíčková et al., 2017; Kramann et al., 2018). More importantly, we also detected an increased PT sub-population, which we characterized as injured PT, by increased accessibility around the injury markers *Vcam1* and *Kim1*(*Havrc1*) (Vaidya et al., 2006) (Appendix Figure A.8).

<sup>1</sup>Prof. Rafael Kramann and Christoph Kuppe supported the annotation.



**Figure 5.16: Visualization of Cicero predicted DNA regulatory elements interactions.** Visualization of co-accessibility scores (y-axis) of Cicero predicted with raw and scOpen estimated matrices contrasted with scores based on RNA pol-II ChIA-PET (purple) and promoter capture Hi-C (green) around the CD79A locus (x-axis). For ChIA-PET, the log-transformed frequencies of each interaction PET cluster represent co-accessibility scores, while the negative log-transformed p-values from the CHiCAGO software indicate Hi-C scores. The ChIA-PET and Hi-C are used as true labels.

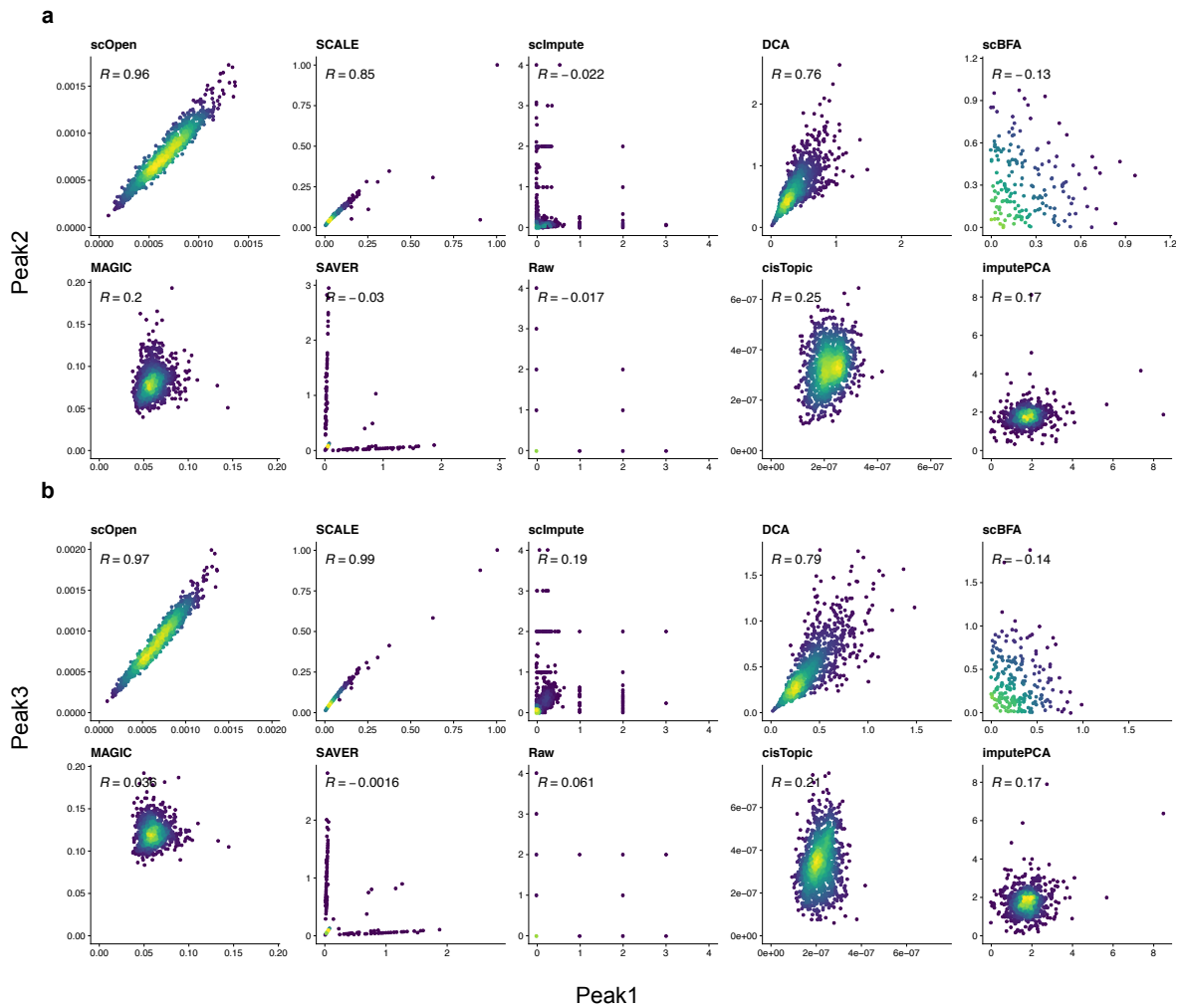
### Estimation of Cell-type-specific TF Activity

Next, we adapted HINT-ATAC (Li et al., 2019), a computational methods for detecting transcription factor binding sites from ATAC-seq (see Section 2.3.1: Computational Footprinting Analysis), to dissect regulatory changes in scATAC-seq clusters. For each cluster, we created a pseudo-bulk ATAC-seq library by combining reads from single cells in the cluster. We then performed footprinting analysis and estimated TF activity scores for all footprint-supported motifs. We only kept TFs with changes (high variance) in TF activity scores among clusters. We focused here on clusters associated with proximal tubular cells (PT), fibroblasts, and immune cells, as these represent key players in kidney remodeling and fibrosis after injury.

As shown in Figure 5.20, the TF activity scores captured regulatory programs associated with these three major cell populations <sup>1</sup>. Interestingly, injured PTs showed overall lower TF activity scores of all TFs of the PT cluster. TFs with a high decrease in activity in injured PTs included *Rxra*, which was known to be important for the regulation of calcium homeostasis in tubular cells (Sugawara et al., 1997), and *Hnf4a*, which is important in proximal tubular development (Marable et al., 2018) (Figure 5.20). Footprint profile of *Rxra* in injured PTs displayed a gradual loss of TF activity over time, indicating that injured PT acquires a de-differentiated phenotype during fibrosis progression and tubular dilatation (Figure 5.20c). A group of TFs with high activity scores in injured PTs also had increased

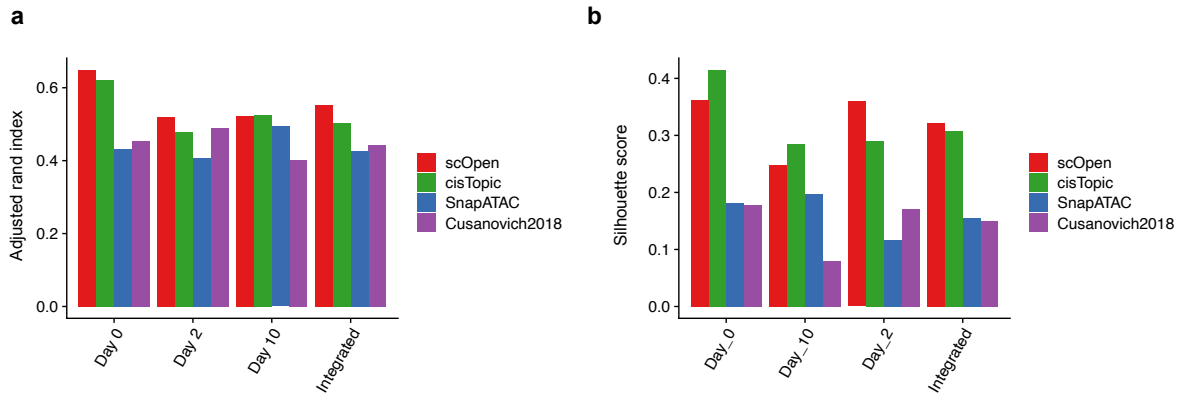
<sup>1</sup>Prof. Rafael Kramann and Christoph Kuppe provided help to interpret the data.

## 5.2. Biological Validation



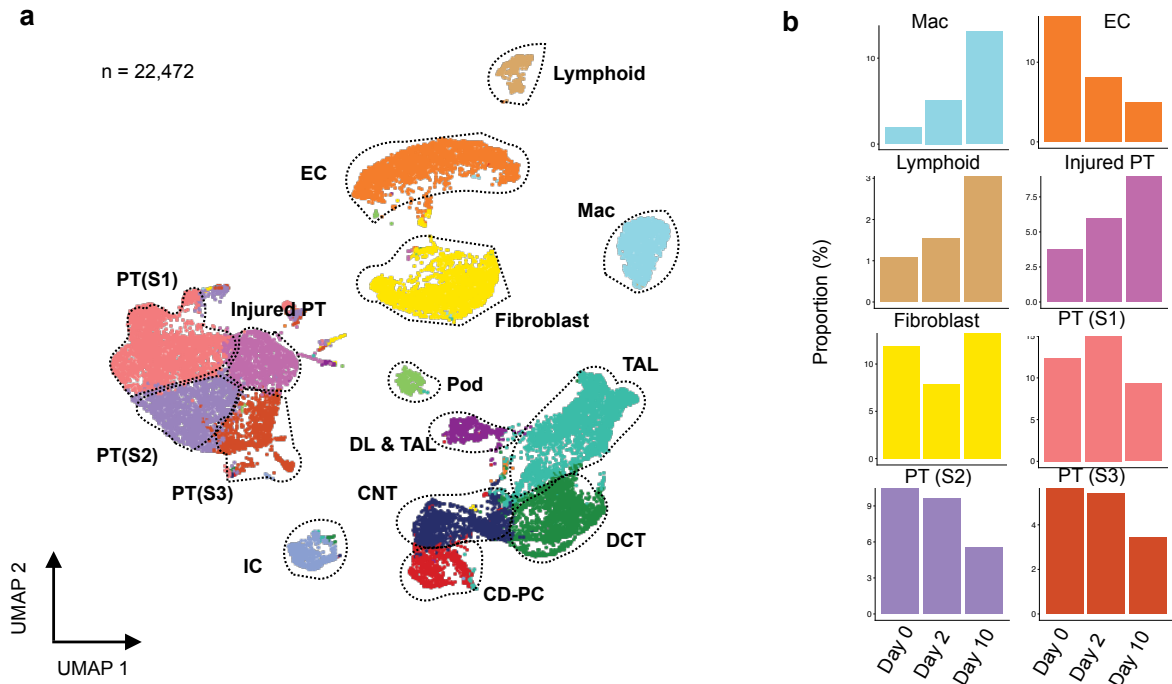
**Figure 5.17: Visualization of co-accessibility score a**, Scatter plot showing single-cell accessibility scores estimated by top-performing imputation methods for the link between peak 1 and peak 2 (supported by Hi-C data). Each dot represents a cell and color refers to density. Pearson correlation is shown on the left-upper corner. **b**, Same as **a** for peak 1 and peak3.

TF activity scores in fibroblasts (Smad2:Smad3 and Batf:Jun), indicating shared regulatory programs in these cells. Smad proteins are downstream mediators of TGF $\beta$  signaling, which is a known key player of fibroblast to myofibroblast differentiation and fibrosis (Kramann et al., 2013). The high activity of Smad2::Smad3 also indicated a role of TGF $\beta$  in the de-differentiation of injured PTs. Interestingly, both Smad2:Smad3 reached a peak in TF activity level at day 2 after UUO in injured PTs (Figure 5.20c), which indicated these TFs are activated post-transcriptionally. Furthermore, we also detect the high activity of Nfkb1 in injured PTs (and lymphocytes), which fits with the known role of Nfkb1 in injured / failed repair PTs (Kirita et al., 2020; Markó et al., 2016). Moreover, our analysis also showed a gradual TF activity increase over time in injured PT (Figure 5.20c), suggesting that Nfkb1 plays an important role in sustaining the injured PT phenotype.



**Figure 5.18: Evaluation of batch correction using different dimensionality reduction methods.**

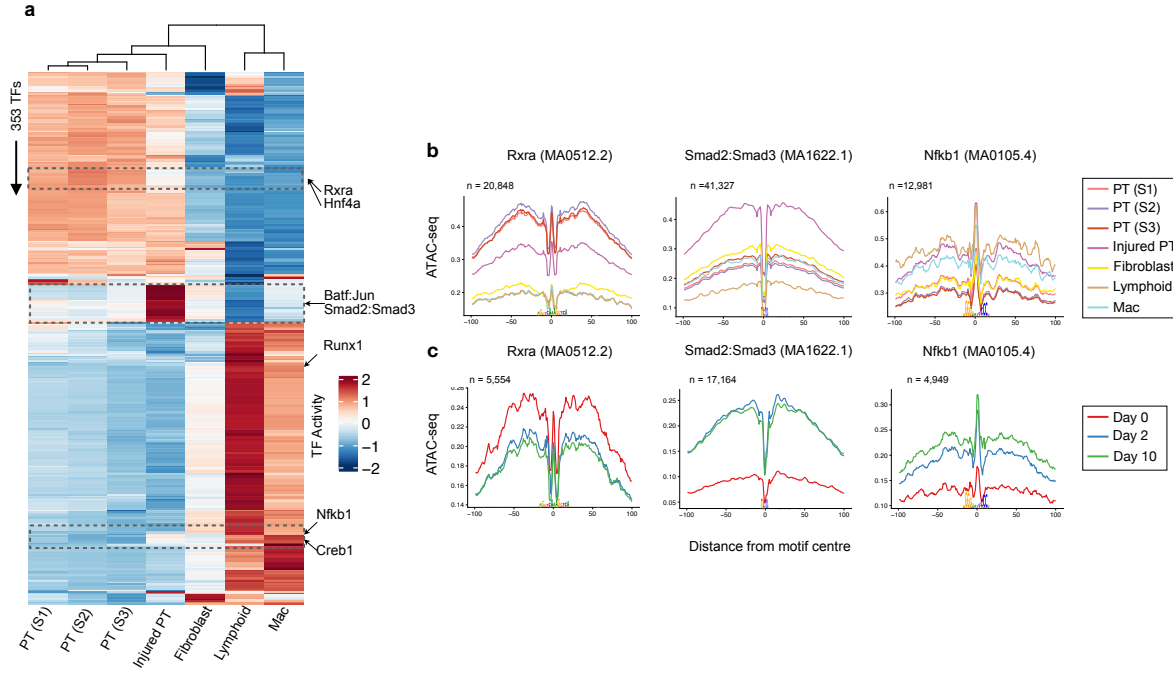
**a**, ARI values (y-axis) contrasting clustering results and transferred labels using distinct dimensional reduction methods for scATAC-seq. Clustering was performed by only considering UWO kidney cells on day 0 (WT), day 2, day 10, and the integrated data set (all samples). **b**, Same as **a** for silhouette score.



**Figure 5.19: Annotation of UWO scATAC-seq data.** **a**, UMAP of the integrated UWO scATAC-seq after doublet removal with major kidney cell types: fibroblasts, descending loop of Henle and thin ascending loop of Henle (DL & TAL); macrophages (Mac), Lymphoid (T and B cells), endothelial cells (EC), thick ascending loop of Henle (TAL), distal convoluted tubule (DCT), collecting duct-principal cells (CD-PC), intercalated cells (IC), podocytes (Pod) and proximal tubule cells (PT S1; PT S2; PT S3; Injured PT). **b**, Proportion of cells of selected clusters on either day 0, day 2, or day 10 experiments.

### 5.2.2 Characterizing Gene Regulation During Myofibroblast Differentiation

## 5.2. Biological Validation

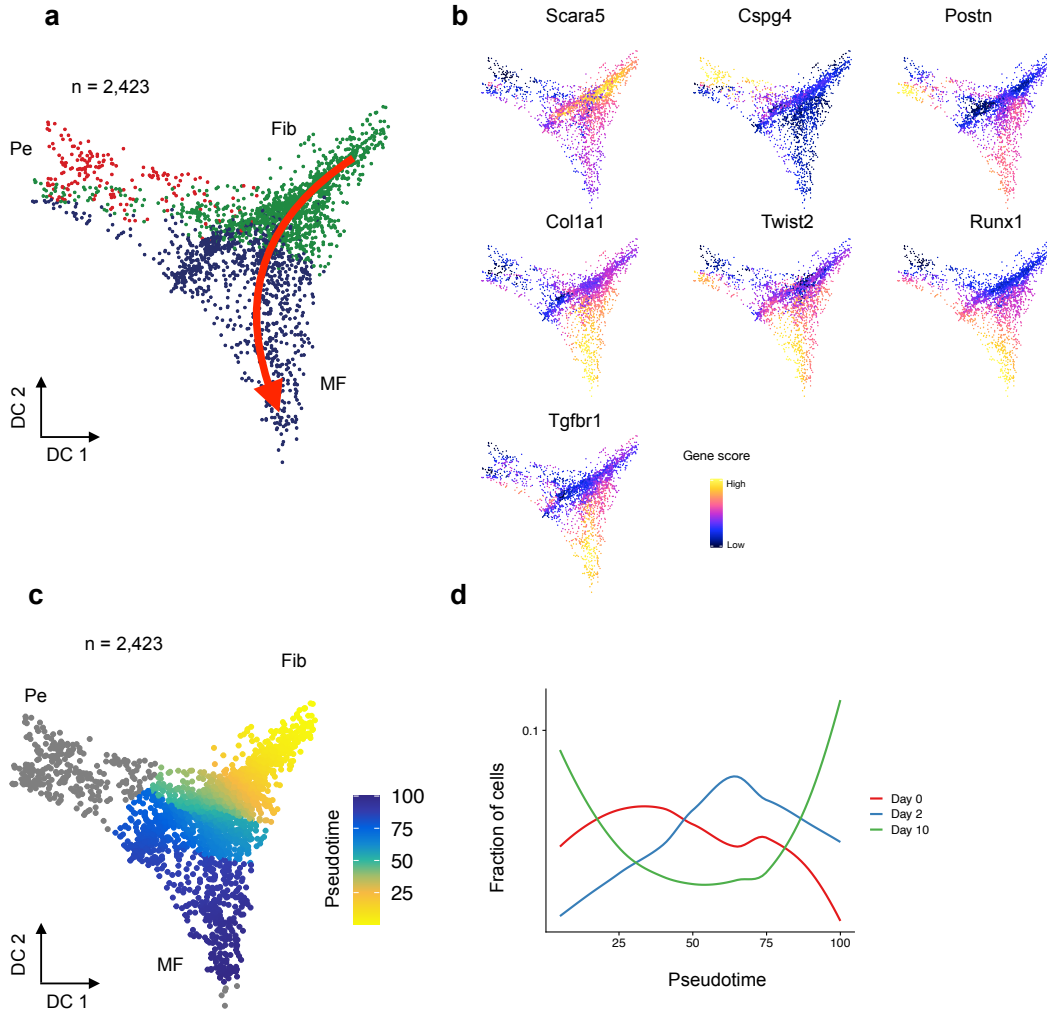


**Figure 5.20: Visualization of TF activity across different cell types and time points. a,** Heatmap with TF activity score (z-transformed) for TFs (y-axis) and selected clusters (x-axis). We highlight TFs with the decrease in activity scores in injured PTs (Rxra and Hnf4a), with high TF activity scores in injured PTs (Batf:Jun; Smad2:Smad3) and immune cells (Creb1; Nfkb1). **b,** Transcription factor footprints (average ATAC-seq around predicted binding sites) of Rxra, Smad2::Smad3 and Nfkb1 for selected cell types. The logo of underlying sequences is shown below and the number of binding sites is shown top-left corner. **c,** Transcription factor footprints of Rxra, Smad2::Smad3 and Nfkb1 for injured PT cells in day 0, day 2 and day 10.

### Identification and Validation of TFs for Myofibroblast Differentiation

A key process following kidney injury is fibrosis, which is caused by the differentiation of fibroblasts and pericytes to matrix secreting myofibroblasts (Kuppe et al., 2021). To dissect potential differentiation trajectories, we performed a diffusion map embedding of the fibroblasts (Figure 5.21a), which revealed the presence of three major branches formed by fibroblasts, pericytes, and myofibroblasts, as supported by the expression of Scara5, Ng2(Cspg4), Postn and Col1a1 (Figure 5.21b) (Kuppe et al., 2021; Muhl et al., 2020). We next created a cellular trajectory across the differentiation from fibroblasts to myofibroblasts using ArchR (Figure 5.21c). We observed that there is an increase in cells after injury (Day 2 and Day 10) along the trajectory (Figure 5.21d).

To identify the driving regulators for this differentiation process, we characterized TFs by correlating their gene activity inferred by ArchR (Section 2.3.2) with TF activity estimated by chromVAR along the trajectory (Figure 5.22a) and ranked these by their correlation (Figure 5.22b). The correlation of Runx1, which has a well-known function in blood cells (de Bruijn and Dzierzak, 2017), stood out, besides showing a steady increase in activity in myofibroblasts. Another TF with high correlation and similar myofibroblast specific activity was Twist2, which has a known role in epithelial to



**Figure 5.21: Visualization of myofibroblast differentiation.** **a**, Diffusion map showing sub-clustering of fibroblasts. Colors refer to sub-cell-types and arrow represents differentiation trajectory from fibroblast to myofibroblast. Pe (pericyte), Fib (fibroblast), MF (myofibroblast). **b**, Scatter plot showing gene activity score for Scara5 (fibroblast), Cspg4 (pericytes), Postn (myofibroblasts), Col1a1 (myofibroblasts), Twist2, Runx1 and Tgfb1. **c**, Diffusion map showing the inferred pseudotime for each cell along the differentiation trajectory. **d**, Fraction of cell for each sample along the trajectory.

mesenchymal transition in kidney fibrosis (Chan et al., 2018) (Figure 5.22c).

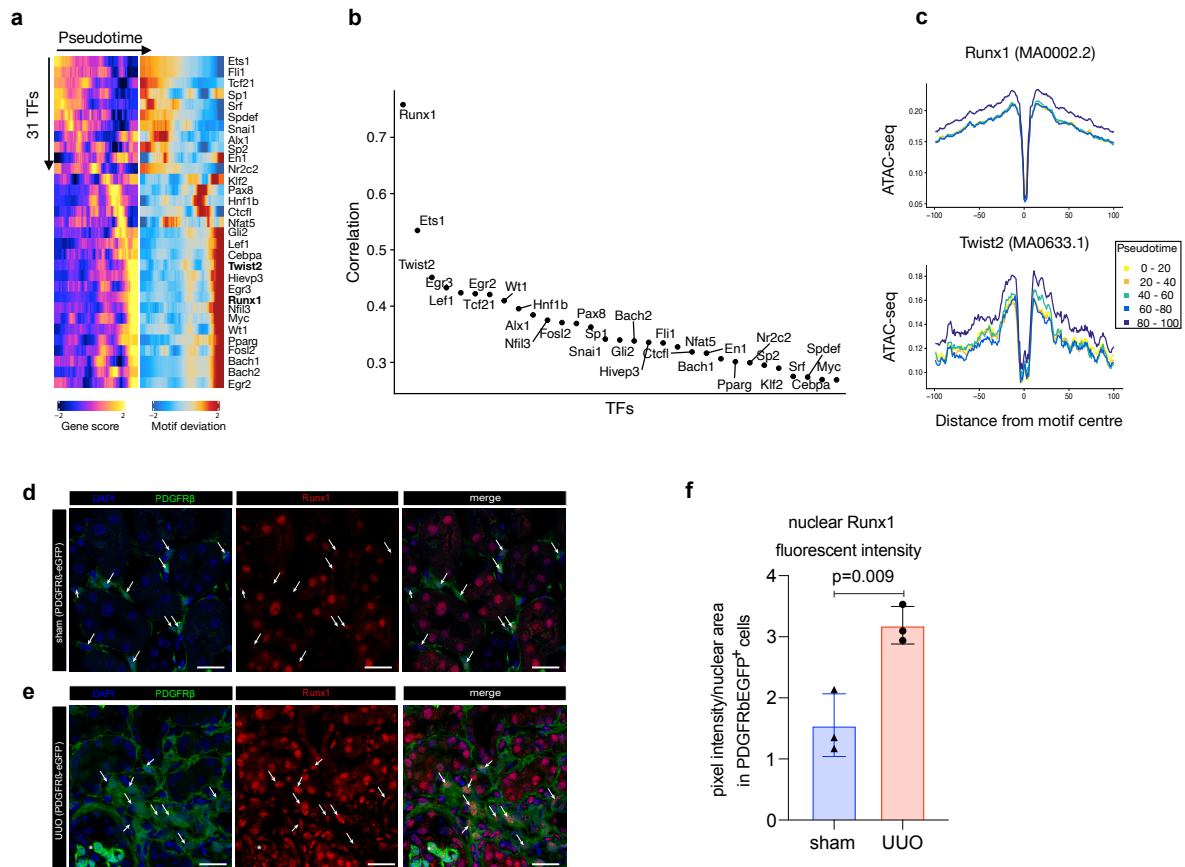
To validate the yet uncharacterized role of Runx1 in myofibroblasts, we performed immunostaining and quantification of Runx1 signal intensity in transgenic PDGFRb-eGFP mice that genetically tag fibroblasts and myofibroblasts (Kuppe et al., 2021)<sup>1</sup>. Runx1 staining in control mice (sham) revealed positive nuclei in tubular epithelial cells and rarely in PDGFRbeGFP+ mesenchymal cells (Figure 5.22d). In kidney fibrosis after UUO surgery (day 10), Runx1 staining intensity increased significantly in PDGFRb+ myofibroblasts (Figure 5.22e-f).

Next, we performed lentiviral overexpression experiments and RNA-sequencing in a human kidney PDGFRb+ fibroblast cell-line that we have generated (Kuppe et al., 2021) to ask whether Runx1

<sup>1</sup>The experiment was performed by Christoph Kuppe.



## 5.2. Biological Validation



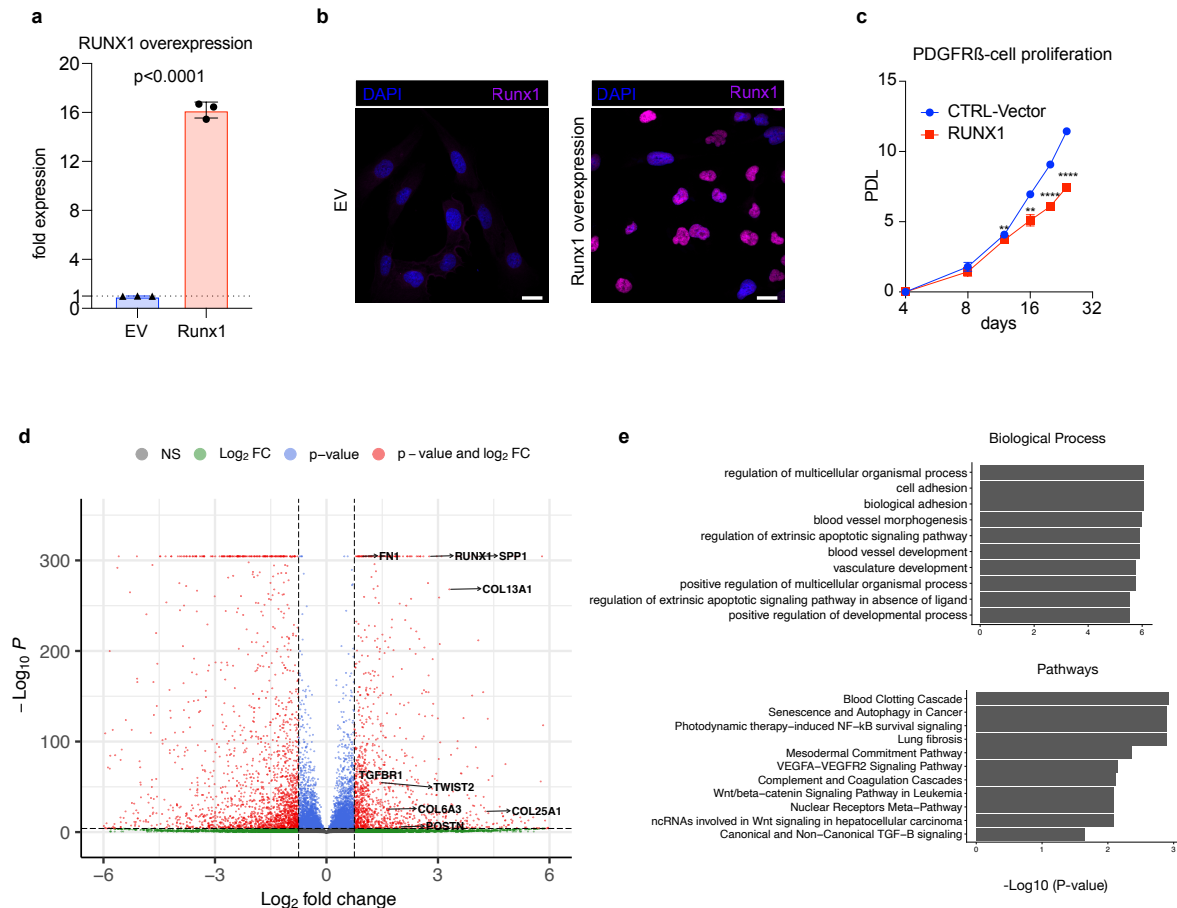
**Figure 5.22: Identification and validation of Runx1 for myofibroblast differentiation.** **a**, Pseudotime heatmap showing gene activity (left) and TF motif activity (right) along the trajectory. **b**, Correlation of gene activity score and motif deviation along the trajectory for selected TFs. Each dot represents a TF and TFs are sorted by their correlation. Runx1 has the highest correlation, followed by ETS1 and Twist2. Of these, Runx1 and Twist2 have an increase in gene scores and TF activity. **c**, Footprinting profiles of Runx1 and Twist2 binding sites along the trajectory. **d**, Immuno-fluorescence (IF) staining of Runx1 (red) in PDGFRb-eGFP mouse kidney. In sham-operated mice Runx1 staining shows a reduced intensity in PDGFRb-eGFP+ cells compared to remaining kidney cells (arrows). **e**, Immuno-fluorescence (IF) staining of Runx1 (red) in PDGFRb-eGFP mouse kidney at 10 days after UUO as compared to sham. Arrows indicate Runx1 staining in expanding PDGFRb-eGFP+ myofibroblasts. **f**, Quantification of Runx1 nuclear intensity in PDGFRb-eGFP+ cells in sham vs. UUO mice. Error bars represent the SD of the intensity. Data are presented as mean  $\pm$  SD. Statistical significance was assessed by a two-tailed Student's t-test with  $p < 0.05$  being considered statistically significant ( $n=3$  mice). Scale bars in **d** and **e** represent 50  $\mu$ m.

might be functionally involved in myofibroblast differentiation in humans (Figure 5.23a-b)<sup>1</sup>. Runx1 overexpression led to reduced proliferation (Figure 5.23c) and strong gene expression changes (Figure 5.23d). Various extracellular matrix genes (Fn1, Col13A1) as well as a TGF $\beta$  receptor (Tgfr1)

<sup>1</sup>Nazanin Kabgani performed the cell culture experiments together with Susanne Ziegler. Nazanin Kabgani also generated the PDGFRb cell line and Susanne Ziegler performed the cloning for Runx1 overexpression. The RNA isolation and bulk RNA Seq library prep was performed by Christoph Kuppe.



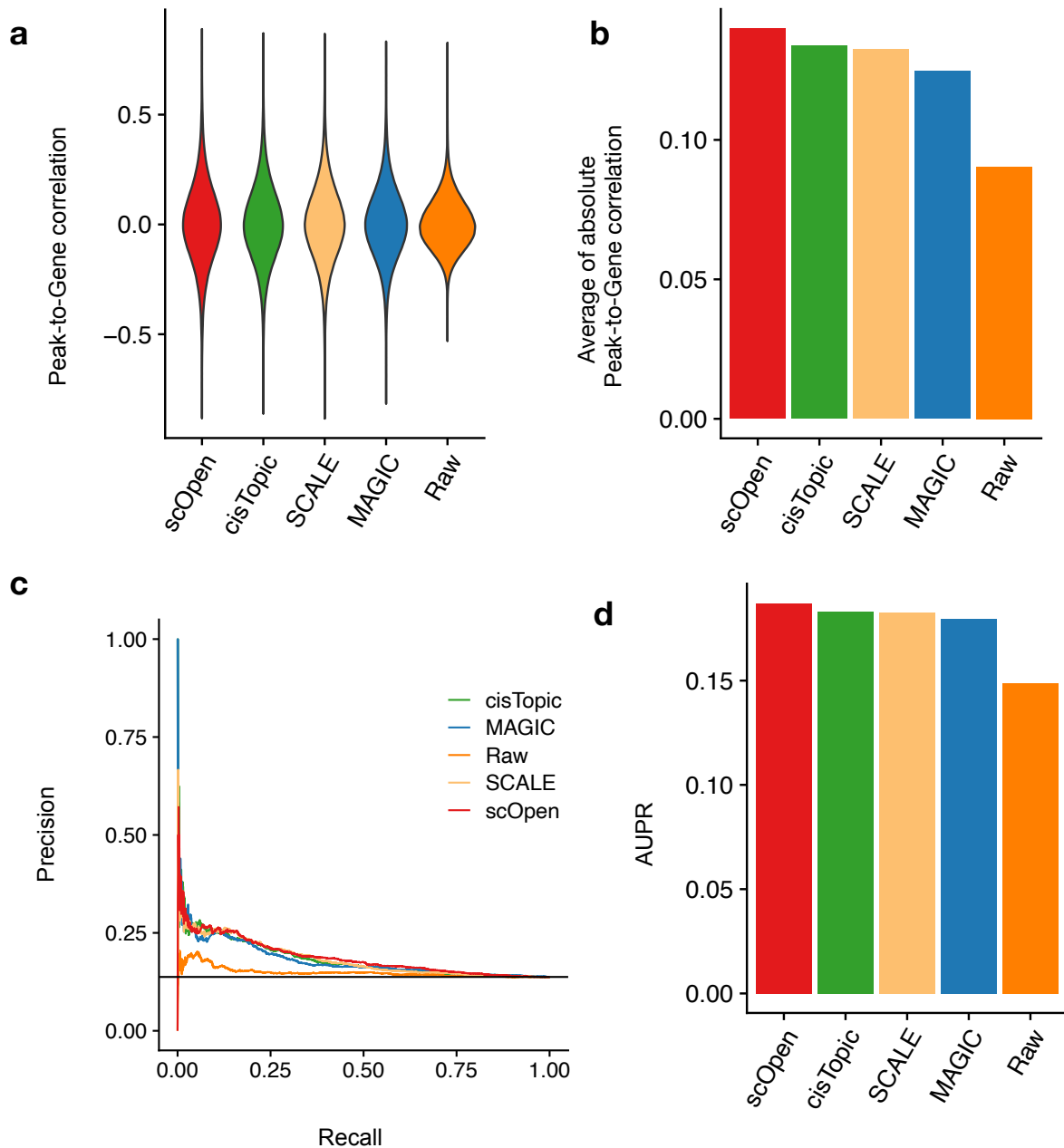
## 5.2. Biological Validation



**Figure 5.23: Overexpression of Runx1.** **a**, Expression of Runx1 by qPCR after lentiviral Runx1 overexpression in a human kidney PDGFR $\beta$ + cell line. Error bars represent the SD of the intensity. Data are presented as mean  $\pm$  SD. Statistical significance was assessed by a two-tailed Student's t-test with  $p < 0.05$  being considered statistically significant ( $n=3$ ). **b**, Immuno-fluorescence (IF) staining of Runx1 in human kidney PDGFR $\beta$ + cells lentivirally transduced with either empty vector (EV) or an Runx1 overexpression construct (Runx1). Scale bars represent 10  $\mu$ m. **c**, Population doubling of Runx1 over-expressing cells vs. control (EV). Statistical significance was assessed by a two-tailed Student's t-test with  $p < 0.05$  being considered statistically significant ( $n=3$ ). **d**, Volcano plot showing differential expression analysis between Runx1 overexpression vs. control. Each dot represents a gene and dashed lines represent the thresholds ( $x = 1.5$  and  $y = 5$ ) used for selection of DE genes. Colors refer to significance given different thresholds. **e**, Barplot showing GO enrichment using up-regulated genes from Runx1 overexpression. Top 10 terms are shown for Biological Process and WikiPathways.

and Twist2 were up-regulated following Runx1 overexpression. GO and pathway enrichment analysis indicated enrichment of cell adhesion, cell differentiation, and TGF $\beta$  signaling following Runx1 overexpression (Figure 5.23e). Furthermore, we observed increased expression of the myofibroblast marker gene Postn after Runx1 overexpression. Altogether, this suggests that Runx1 might directly drive myofibroblast differentiation of human kidney fibroblasts since overexpression reduced cell proliferation and induced expression of various myofibroblast genes.

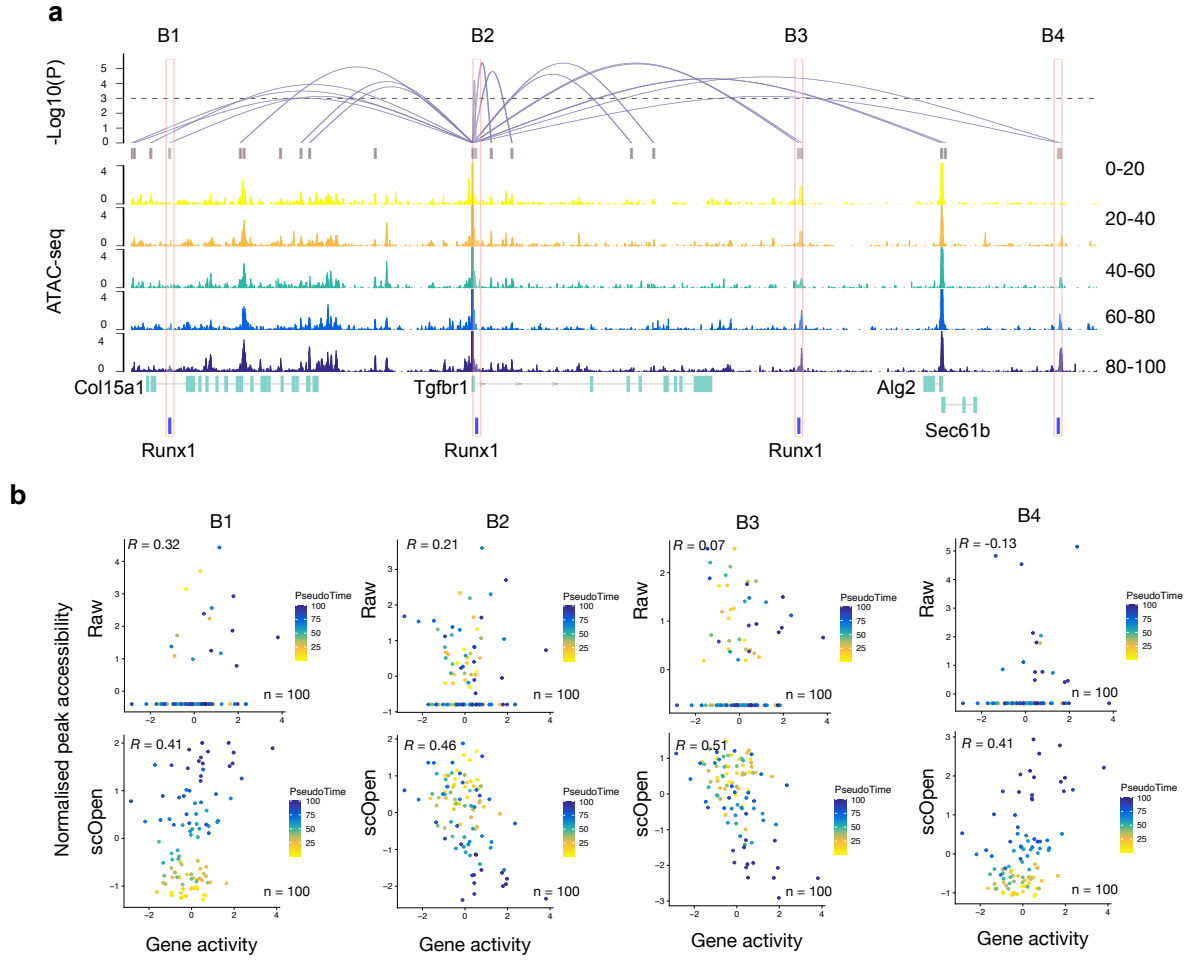
## 5.2. Biological Validation



**Figure 5.24: Prediction of Runx1 target genes.** **a**, Violin plot showing predicted Peak-to-Gene (P2G) links using raw or imputed matrix from scOpen, cisTopic, SCALE and MAGIC. Methods are sorted by average absolute value. **b**, Average absolute value of correlation of P2G links. **c**, Precision-Recall curves of predicted Runx1 target genes using raw, scOpen, SCALE, MAGIC or cisTopic matrix. **d**, Barplot showing Performance of top-performing imputation methods on the prediction of Runx1 target genes as measured by AUPR.

### 5.2.3 Identification and Validation of Runx1 Target Genes

Another important application of scATAC-seq is the prediction of *cis*-regulatory DNA interactions (i.e., co-accessibility analysis) by measuring the correlation between gene activity and reads counts



**Figure 5.25: Tgfb1 is regulated by Runx1 in myofibroblast.** **a** Peak-to-Gene links (top) predicted on scOpen matrix and associated to Tgfb1 in fibroblast cells. The height of links represents its significance. Dash line represents the threshold of significance ( $FDR = 0.001$ ). ATAC-seq tracks (below) were generated from pseudo-bulk profiles of fibroblast/myofibroblast cells with increasing pseudo time (0-20, 20-40, 40-60, 60-80, and 80-100). Binding sites of Runx1 (B1-B4) supported by ATAC-seq footprints and overlapping to peaks are highlighted on the bottom. **b** Scatter plot showing gene activity of Tgfb1 and normalized peak accessibility from raw (upper) or scOpen imputed matrix (lower) for peak-to-gene link B1, B2, B3 and B4. Each dot represents cells in a given pseudotime and the overall correlation is shown in the left-upper corner.

in proximal peaks (see Section 2.3.2). As shown in Section 5.1.5, we have demonstrated that this prediction can be significantly improved by using the imputed matrix. Here, we predicted the peak-to-gene links in fibroblasts on distinct scATAC-seq matrices after imputation with the top-performing imputation methods (i.e., scOpen, cisTopic, SCALE and MAGIC). As expected, the use of imputation methods led to improved signals on peak-to-gene links predictions as indicated by higher correlation values after imputation compared with the correlation using raw matrix (Figure 5.24a-b).

We next sought to identify the target genes of Runx1 by leveraging the predicted peak-to-gene links. For this, we considered all genes with at least one link, where the peak has a footprint sup-

### 5.3. Discussion

ported Runx1 binding site, as Runx1 targets. We then compared the predicted Runx1 targets from distinct scATAC-seq imputed matrices with differentially expressed genes after Runx1 overexpression (true labels). Interestingly, all imputation methods obtained higher AUPR values than the use of a raw matrix, while scOpen obtained the highest AUPR (Figure 5.24c-d). Among others, scOpen predicted *Tgfbr1* and *Twist2* as prominent Runx1 target genes (Figure 5.25a; Appendix Figure A.10a). We observed several peaks with high peak-to-gene correlation, increasing accessibility upon myofibroblast differentiation and presence of Runx1 binding sites. The positive impact of imputation was clear when observing scatter plots contrasting gene activity and peak accessibility of these peak-to-gene links (Figure 5.25b; Appendix Figure A.10b).

These results suggest that Runx1 is an important regulator of myofibroblast differentiation by regulating the EMT-related TF *Twist2* and by increasing the expression of a TGF $\beta$  receptor to amplify TGF $\beta$  signaling by and affect the expression of extracellular matrix genes. Altogether, these results uncover a complex cascade of regulatory events across cells during the progression of fibrosis and reveal a yet unknown function of Runx1 in myofibroblast differentiation in kidney fibrosis.

## 5.3 Discussion

---

In this chapter, we presented all the benchmarking results obtained by performed the experiments described in Chapter 4. For technical validation, we evaluated the performance of our proposed method, scOpen, and showed the outperformance of scOpen compared with the competing methods from different perspectives. For biological validation, we applied scOpen to a complex scATAC-seq data generated from the intact mouse kidneys at different time points after injuring. Our analyses indicated that scOpen can provide better integration results and is able to identify all major cell types for mouse kidney, as evaluated using an independent snRNA-seq from the same via label transferring approach (Stuart et al., 2019). Moreover, we identified and validated Runx1 as an important regulator for myofibroblast differentiation, thus revealing novel biological insights for the fibrosis process.

---

## Discussion and Conclusion

---

### 6.1 Discussion

---

The sparsity and high-dimensionality in scATAC-seq pose serious challenges to the computation community. The main goal of this work was to develop an algorithm that addresses these two issues. For this, we derived scOpen, an efficient and effective computational method for single-cell open chromatin data analysis via non-negative matrix factorization (NMF) modeling. scOpen is a linear model that performs matrix imputation and dimensionality reduction on scATAC-seq data. By taking the raw count matrix as input, it first normalizes the data to correct potential technical effects by binarization and term frequency-inverse document frequency (TF-IDF) transformation. Elements in the normalized matrix typically reflect the importance of a peak to a particular cell. Next, it factorizes the normalized matrix into two low-rank matrices (i.e., factors) by using regularized NMF technique. The estimated factors are used as dimension-reduced matrices for either cells or peaks. Furthermore, the multiplication of these two factors is regarded as an imputed or denoised matrix which is used for downstream analysis.

Two parameter, i.e., the number of components  $k$  and regularization  $\lambda$ , are introduced in this step (see Section 3.3.4). To determine the parameter  $k$ , we introduced a computational algorithm to automatically estimate the number of components for NMF model by detecting the elbow point of reconstruction error against the rank of factors. We showed that in an example scATAC-seq dataset, this approach is able to identify the underlying dimensions, i.e., the number of cell types presented in the data. To further investigate the influence of the hyper-parameter on imputation and dimensionality reduction, we simulated a scATAC-seq dataset by using a novel strategy (Section 4.1.1) and evaluated the performance of scOpen given different parameters in Section 5.1.1.

### Comprehensive Evaluation of Computational Methods for scATAC-seq Analysis

We performed an in-depth evaluation to benchmark a number of computational methods for scATAC-seq data imputation, dimensionality reduction, and downstream analysis. In Section 5.1.2, we demonstrated here that scOpen estimated matrices have a higher recovery of dropout events and also improved distance and clustering results when compared to imputation methods developed for scRNA-seq (Van Dijk et al., 2018; Huang et al., 2018; Li and Li, 2018; Eraslan et al., 2019; Li and Quon, 2019) and the few available imputation methods tailored for scATAC-seq (cisTopic-impute González-Blas et al. (2019), SCALE (Xiong et al., 2019)). scOpen also presented very good scalability with the

## 6.1. Discussion

lowest memory requirements and tractable computational time on large data sets. From a methodological perspective, scOpen is one of the two methods (another one is scBFA) that regularize the models to prevent over-fitting. This is in line with a previous study, which indicated over-fitting as one of the largest issues on scRNA-seq imputation (Andrews and Hemberg, 2018). Moreover, it is also possible to use the scOpen factorized matrix as a dimension reduction. We have shown that both dimensions reduced and imputed matrices from scOpen displayed the best performance on distance representation and clustering when compared to diverse state-of-the-art scATAC-seq dimension reduction/clustering pipelines (cisTopic, SnapATAC and Cusanovich2018) in Section 5.1.3

Finally, we demonstrated that the use of scOpen imputed matrices improves the accuracy of existing state-of-art scATAC-seq methods (cisTopic (González-Blas et al., 2019), chromVAR (Schep et al., 2017), Cicero (Pliner et al., 2018)) in Section 5.1.4. Particularly positive results were obtained in the prediction of chromatin conformation with Cicero, where all methods perform better than raw matrices. Cicero works by measuring the correlation between pairs of proximal links. Due to the fact that dropout events are independent for two regions, it is not surprising that imputation has strong benefits. This is equivalent to observations from Van Dijk et al. (2018) in the context of scRNA-seq, where the prediction of gene-gene interactions after MAGIC imputation was significantly improved. Altogether, these results support the importance of dropout event correction with scOpen in any computational analysis of scATAC-seq (Section 5.1.5).

### Gaining Novel Biological Insights with scOpen

For biological validation, we used scOpen to characterize complex cascades of regulatory changes associated with kidney injury and fibrosis in Section 5.2.1. Our analyses demonstrated that a major expanding population of cells, i.e. injured PTs, myofibroblasts, and immune cells, share regulatory programs, which are associated with cell de-/differentiation and proliferation. Of all methods evaluated, scOpen obtained the best clustering results in the kidney cell repertoire using a scRNA-seq on the same kidney injury model as a reference (Figure 5.18). Moreover, we estimated cell-type-specific TF binding activity by computational footprinting analysis using HINT-ATAC (Li et al., 2019) as described in Section 2.3.1 and identified a number of specific TFs for the relevant cell types.

The differentiation from fibroblasts to myofibroblast, also known as fibrosis, plays a key role in kidney injury (Kuppe et al., 2021). To understand the gene regulatory mechanisms, we further analyzed the fibroblast cells (Section 5.2.2). We identified Runx1 as the major TF driving myofibroblast differentiation by trajectory analysis. This was validated by Runx1 staining in the mouse model and by lentiviral over-expression studies in human PDGFRb+ kidney cells. Computational prediction with peak-to-gene links combined with footprint-supported Runx1 binding sites indicated the role of Runx1 in the regulation of *Tgfb1* and *Twist2*. These were validated on over-expression experiments in human fibroblasts. Altogether, the results suggested that Runx1 makes fibroblasts more sensitive to TGFB signaling via increasing expression of the TGFB receptors. Here, we showed for the first time in-vivo and in-vitro evidence that Runx1 in myofibroblasts regulates scar formation following a fibrogenic kidney injury in mice. Runx1 deficiency caused reduced myofibroblast formation and enhanced recovery. To this end, inhibiting Runx1 could lead to reduced myofibroblast differentiation

and increased endogenous repair after fibrogenic organ injuries in the kidney and heart. Our results shed light on mechanisms of myofibroblasts differentiation driving kidney fibrosis and chronic kidney disease (CKD). Altogether, this demonstrated how scOpen can be used to dissect complex regulatory processes by footprinting analysis combined with peak-to-gene link predictions.

## 6.2 Conclusion and Future Work

---

### Improving the Scalability of scOpen

With the advancement of single-cell sequencing technology, scATAC-seq data continues to grow at an unprecedented pace and atlas-scale datasets have been rapidly generated. For example, In one study, Silvia et al. (2020) profiled the chromatin accessibility of nearly 800,000 cells from 59 human samples across 15 organs. In another study, Zhang et al. (2021) applied scATAC-seq to 25 adult human tissues, generating the open chromatin profiles for roughly 500,000 cells. Although we have demonstrated that scOpen has better scalability compared with the competitors in Section 5.1.2 (Figure 5.4), the largest benchmarking dataset (i.e., *PBMC*) contains only about 10,000 cells (see Table 4.1), representing a relatively small dataset. Therefore, it is interesting to further improve the scalability of scOpen. To reduce the memory usage, a promising approach is to use online machine learning for non-negative matrix factorization in which only a subset of the data is used to update the parameters at each step of the algorithm (Mairal et al., 2010; Lefevre et al., 2011). For example, the online NMF has been shown to be able to iteratively and incrementally integrate more than 1 million cells using about 1.9GB of RAM (Gao et al., 2021). To speed scOpen up, one can seek to employ a graphics processing unit (GPU) to solve the NMF optimization problem by taking advantages of the high computing performance delivered by GPU (Mejía-Roa et al., 2015). We envision that incorporating the online learning into scOpen and implementing the NMF algorithm based on GPU would significantly extend its application to large-scale scATAC-seq data.

### Applying scOpen to Other Single-cell Protocols

In this thesis, we applied scOpen to single-cell chromatin accessibility data as measured by scATAC-seq protocol. On one hand, we have demonstrated that scOpen is able to produce a better imputed matrix and dimension reduced matrix compared with the competing methods. On the other hand, the development of technology has allowed for profiling other chromatin features at single-cell resolution, such as histone modification by scChIP-seq (Rotem et al., 2015; Bartosovic et al., 2021), transcription factor binding sites by scCUT&Tag (Kaya-Okur et al., 2019), and DNA methylation by scBisulfite-seq (Smallwood et al., 2014). Although these protocols aim at measuring different type of chromatin states (i.e., open versus closed in scATAC-seq, modified versus unmodified in scChIP-seq, bound versus unbound in scCUT&Tag, and methylated versus unmethylated in scBisulfite-seq), the data generated by them share similar properties, i.e., sparsity, high-dimensionality, and low count information. Therefore, it is interesting to use the same methodology from scOpen to perform imputation and dimensionality reduction for these protocols.

### Exploiting the Low-dimensional Matrix for Peaks

In addition to generate a dimension reduced matrix for cells (i.e.,  $\mathbf{H}$ ), scOpen produces a low-dimensional representation for peaks via NMF (i.e.,  $\mathbf{W}$ , see Section 3.3.1 and Figure 3.2). In this thesis, we showed that the matrix  $\mathbf{H}$  can be used for cell clustering, integration and visualization (Section 5.1.3). Similar to scOpen, cisTopic also generates a low-dimensional matrix containing *cis*-regulatory topics (see Section 2.4.3). Furthermore, González-Blas et al. (2019) demonstrated that the topics can be used for motif discovery to predict transcription factors and to explore chromatin state variations. Therefore, how to exploit the low-dimensional matrix for peaks as generated by scOpen represents a future challenge.

### Extending scOpen to Single-cell Multimodal Omics Data Integration

Advances in single-cell genomics technologies have enabled measurement gene expression and chromatin accessibility simultaneously in the same cell, such as Paired-seq (Zhu et al., 2019) and SHARE-seq (Ma et al., 2020). A key challenge in analyzing the single-cell multimodal omics data is how to perform cross-modal integration for downstream analysis (e.g., clustering and visualization) while controlling confounding factors (e.g., batch effects), given that the data from each modality has characteristic statistical, technical and biological features. This problem can be formalized as multi-view learning where each modality is considered as a view for an individual cell. Interestingly, several works have demonstrated that NMF is a powerful and flexible model for addressing this issue (Liu et al., 2013, 2014; Zhang et al., 2018; Liang et al., 2020). Therefore, we would like to investigate how to extend scOpen to allow for learning a low-dimensional representation from single-cell multimodal data efficiently and effectively in the future.



---

## Bibliography

---

- Alberts B., Johnson A., Lewis J., et al. *Molecular Biology of the Cell*. Garland Science Publ., Sixth edition, 2017.
- Andrews T. S. and Hemberg M. False signals induced by single-cell imputation. *F1000Research*, 7, 2018.
- Angerer P., Haghverdi L., Büttner M., et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–1243, apr 2016.
- Bábíčková J., Klinkhammer B. M., Buhl E. M., et al. Regardless of etiology, progressive renal disease causes ultrastructural and functional alterations of peritubular capillaries. *Kidney international*, 91(1):70–85, 2017.
- Bartosovic M., Kabbe M., and Castelo-Branco G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nature Biotechnology*, 2021.
- Becht E., McInnes L., Healy J., et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, 2019.
- Benjamini Y. and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, aug 1995.
- Bentsen M., Goymann P., Schultheis H., et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications*, 11(1):4267, 2020.
- Blei D. M., Ng A. Y., and Jordan M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
- Blondel V. D., Guillaume J.-L., Lambiotte R., and Lefebvre E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- Boutsidis C. and Gallopoulos E. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.
- Buenrostro J. D., Giresi P. G., Zaba L. C., Chang H. Y., and Greenleaf W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–8, 2013.

## Bibliography

- Buenrostro J. D., Wu B., Chang H. Y., and Greenleaf W. J. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109(1):21.29.1–21.29.9, 2015a.
- Buenrostro J. D., Wu B., Litzenburger U. M., et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015b.
- Buenrostro J. D., Corces M. R., Lareau C. A., et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6):1535–1548, 2018.
- Cairns J., Freire-Pritchett P., Wingett S. W., et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biology*, 17(1):127, 2016.
- Cao J., Spielmann M., Qiu X., et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.
- Chan S. C., Zhang Y., Shao A., et al. Mechanism of fibrosis in HNF1B-related autosomal dominant tubulointerstitial kidney disease. *Journal of the American Society of Nephrology*, 2018.
- Chen H., Lareau C., Andreani T., et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology*, 20(1):241, 2019.
- Chen X., Miragaia R. J., Natarajan K. N., and Teichmann S. A. A rapid and robust method for single cell chromatin accessibility profiling. *Nature Communications*, 9(1):5345, 2018.
- Cichocki A. and Phan A.-H. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.
- Coifman R. R., Lafon S., Lee A. B., et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005.
- Corces M. R., Buenrostro J. D., Wu B., et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*, 48(10):1193–1203, 2016.
- Corces M. R., Granja J. M., Shams S., et al. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413), 2018.
- Cusanovich D. A., Daza R., Adey A., et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.
- Cusanovich D. A., Hill A. J., Aghamirzaie D., et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5):1309–1324, 2018a.
- Cusanovich D. A., Reddington J. P., Garfield D. A., et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, 555(7697):538–542, 2018b.

- Danese A., Richter M. L., Chaichoompu K., et al. EpiScanpy: integrated single-cell epigenomic analysis. *Nature Communications*, 12(1):5228, 2021.
- Davis J. and Goadrich M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.
- Bruijn M.d and Dzierzak E. Runx transcription factors in the development and function of the definitive hematopoietic system. *Blood*, 129(15):2061–2069, 2017.
- De-Leon S. B.-T. and Davidson E. H. Gene Regulation: Gene Control Network in Development. *Annual Review of Biophysics and Biomolecular Structure*, 36(1):191–212, may 2007.
- Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., and Harshman R. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- Demšar J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006.
- Dhillon I. S. and Sra S. Generalized nonnegative matrix approximations with bregman divergences. In *NIPS*, volume 18. Citeseer, 2005.
- Dobin A., Davis C. A., Schlesinger F., et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, jan 2013.
- Domcke S., Hill A. J., Daza R. M., et al. A human cell atlas of fetal chromatin accessibility. *Science*, 370(6518), 2020.
- Eraslan G., Simon L. M., Mircea M., Mueller N. S., and Theis F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, 2019.
- Fang R., Preissl S., Li Y., et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature Communications*, 12(1):1337, 2021.
- Fornes O., Castro-Mondragon J. A., Khan A., et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 48(D1):D87–D92, jan 2020.
- Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.
- Friedman M. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.
- Gao C., Liu J., Kriebel A. R., et al. Iterative single-cell multi-omic integration using online learning. *Nature Biotechnology*, 39(8):1000–1007, 2021.
- Gemmeke J. F., Vuegen L., Karsmakers P., Vanrumste B., and others . An exemplar-based nmf approach to audio event detection. In *2013 IEEE workshop on applications of signal processing to audio and acoustics*, pages 1–4. IEEE, 2013.

## Bibliography

- González-Blas C. B., Minnoye L., Papasokrati D., et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods*, 16(5):397–400, 2019.
- Granja J. M., Corces M. R., Pierce S. E., et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3):403–411, 2021.
- Grau J., Grosse I., and Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, 31(15):2595–2597, aug 2015.
- Griffiths T. L. and Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.
- Guillamet D., Vitrià J., and Schiele B. Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recognition Letters*, 24(14):2447–2454, 2003.
- Gusmao E. G., Allhoff M., Zenke M., and Costa I. G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nature methods*, 13(4):303–9, 2016.
- Hashimshony T., Wagner F., Sher N., and Yanai I. Cel-seq: Single-cell rna-seq by multiplexed linear amplification. *Cell Reports*, 2(3):666–673, 2012.
- Huang M., Wang J., Torre E., et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539–542, 2018.
- Hubert L. and Arabie P. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.
- Institute B. Picard Tools. <http://broadinstitute.github.io/picard/>, 2019. Accessed: 2019-01-01; version 2.18.22.
- Islam S., Kjällquist U., Moliner A., et al. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167, 2011.
- Josse J. and Husson F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software, Articles*, 70(1):1–31, 2016.
- Kalhor R., Tjong H., Jayathilaka N., Alber F., and Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(1):90–98, 2012.
- Kaya-Okur H. S., Wu S. J., Codomo C. A., et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1930, dec 2019.
- Kharchenko P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, 18(7):723–732, 2021.
- Kim H. and Park H. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM journal on matrix analysis and applications*, 30(2):713–730, 2008.

- Kirita Y., Wu H., Uchimura K., Wilson P. C., and Humphreys B. D. Cell profiling of mouse acute kidney injury reveals conserved cellular responses to injury. *Proceedings of the National Academy of Sciences*, 117(27):15874–15883, jul 2020.
- Klein A., Mazutis L., Akartuna I., et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.
- Klemm S. L., Shipony Z., and Greenleaf W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.
- Korsunsky I., Millard N., Fan J., et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019.
- Kramann R., DiRocco D. P., and Humphreys B. D. Understanding the origin, activation and regulation of matrix-producing myofibroblasts for treatment of fibrotic disease. *The Journal of pathology*, 231(3):273–289, 2013.
- Kramann R., Schneider R. K., DiRocco D. P., et al. Perivascular Gli1+ progenitors are key contributors to injury-induced organ fibrosis. *Cell stem cell*, 16(1):51–66, 2015.
- Kramann R., Machado F., Wu H., et al. Parabiosis and single-cell RNA sequencing reveal a limited contribution of monocytes to myofibroblasts in kidney fibrosis. *JCI insight*, 3(9), 2018.
- Kuppe C., Ibrahim M. M., Kranz J., et al. Decoding myofibroblast origins in human kidney fibrosis. *Nature*, 589(7841):281–286, jan 2021.
- Langmead B. and Salzberg S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, apr 2012.
- Lara-Astiaso D., Weiner A., Lorenzo-Vivas E., et al. Chromatin state dynamics during blood formation. *Science*, 345(6199):943–949, 2014.
- Lareau C. A., Duarte F. M., Chew J. G., et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*, 37(8):916–924, 2019.
- Lee D. and Seung H. S. Algorithms for non-negative matrix factorization. In Leen T., Dietterich T., and Tresp V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.
- Lee D. D. and Seung H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- Lefevre A., Bach F., and Févotte C. Online algorithms for nonnegative matrix factorization with the itakura-saito divergence. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 313–316. IEEE, 2011.
- Li H. and Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

## Bibliography

- Li H., Handsaker B., Wysoker A., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009.
- Li R. and Quon G. scBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biology*, 20(1):193, 2019.
- Li W. V. and Li J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, 9(1):997, 2018.
- Li Z., Schulz M. H., Look T., et al. Identification of transcription factor binding sites using ATAC-seq. *Genome Biology*, 20(1):45, 2019.
- Liang N., Yang Z., Li Z., Sun W., and Xie S. Multi-view clustering by non-negative matrix factorization with co-orthogonal constraints. *Knowledge-Based Systems*, 194:105582, 2020.
- Lin C. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*, 19(10):2756–2779, 2007.
- Liu J., Wang C., Gao J., and Han J. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM international conference on data mining*, pages 252–260. SIAM, 2013.
- Liu J., Jiang Y., Li Z., Zhou Z.-H., and Lu H. Partially shared latent factor learning with multiview data. *IEEE transactions on neural networks and learning systems*, 26(6):1233–1246, 2014.
- Love M. I., Huber W., and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.
- Ma S., Zhang B., LaFave L. M., et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 183(4):1103–1116.e20, 2020.
- Mairal J., Bach F., Ponce J., and Sapiro G. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1), 2010.
- Mann H. B. and Whitney D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.
- Marable S. S., Chung E., Adam M., Potter S. S., and Park J.-S. Hnf4a deletion in the mouse kidney phenocopies Fanconi renal tubular syndrome. *JCI Insight*, 3(14), jul 2018.
- Markó L., Vigolo E., Hinze C., et al. Tubular Epithelial NF- $\kappa$ B Activity Regulates Ischemic AKI. *Journal of the American Society of Nephrology*, 27(9):2658–2669, sep 2016.
- Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17(1):10–12, 2011.
- Mejía-Roa E., Tabas-Madrid D., Setoain J., et al. NMF-mGPU: non-negative matrix factorization on multi-GPU systems. *BMC Bioinformatics*, 16(1):43, 2015.

- Muhl L., Genové G., Leptidis S., et al. Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. *Nature Communications*, 2020.
- Nemenyi P. *Distribution-free Multiple Comparisons*. Princeton University, 1963.
- Paatero P. and Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.
- Patel H., Ewels P., Peltzer A., et al. nf-core/rnaseq: nf-core/rnaseq v3.0 - silver shark, Dec. 2020.
- Patro R., Duggal G., Love M. I., Irizarry R. A., and Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017.
- Patrino L., Maspero D., Craighero F., et al. A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Briefings in Bioinformatics*, 22(4), jul 2021.
- Pierce B. A. *Genetics: A Conceptual Approach*. Macmillan, Fourth edition, 2012.
- Pliner H. A., Packer J. S., McFaline-Figueroa J. L., et al. Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. *Molecular cell*, 71(5):858–871, 2018.
- Rotem A., Ram O., Shores N., et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–1172, 2015.
- Rousseeuw P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- Salton G. and McGill M. J. Introduction to modern information retrieval. 1986.
- Satopaa V., Albrecht J., Irwin D., and Raghavan B. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171, 2011.
- Satpathy A. T., Saligrama N., Buenrostro J. D., et al. Transcript-indexed ATAC-seq for precision immune profiling. *Nature Medicine*, 24(5):580–590, 2018.
- Satpathy A. T., Granja J. M., Yost K. E., et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, 37(8):925–936, 2019.
- Schep A. N., Buenrostro J. D., Denny S. K., et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome research*, 25(11):1757–1770, 2015.
- Schep A. N., Wu B., Buenrostro J. D., and Greenleaf W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, 14(10):975–978, 2017.

## Bibliography

- Shiga M., Seno S., Onizuka M., and Matsuda H. SC-JNMF: Single-cell clustering integrating multiple quantification methods based on joint non-negative matrix factorization. *bioRxiv*, page 2020.09.30.319921, jan 2020.
- Silvia D., J. H. A., M. D. R., et al. A human cell atlas of fetal chromatin accessibility. *Science*, 370(6518):eaba7612, nov 2020.
- Smallwood S. A., Lee H. J., Angermueller C., et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820, aug 2014.
- Stuart T., Butler A., Hoffman P., et al. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.
- Stuart T., Srivastava A., Lareau C., and Satija R. Multimodal single-cell chromatin analysis with Signac. *bioRxiv*, page 2020.11.09.373613, jan 2020.
- Sugawara A., Sanno N., Takahashi N., Osamura R. Y., and Abe K. Retinoid X Receptors in the Kidney: Their Protein Expression and Functional Significance. *Endocrinology*, 138(8):3175–3180, aug 1997.
- Sun S., Zhu J., Ma Y., and Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, 20(1):269, 2019.
- Tang F., Barbacioru C., Wang Y., et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.
- Tarbell E. D. and Liu T. HMMRATAC: a Hidden Markov Modeler for ATAC-seq. *Nucleic Acids Research*, 47(16):e91–e91, sep 2019.
- Taudt A., Nguyen M. A., Heinig M., Johannes F., and Colomé-Tatché M. chromstaR: Tracking combinatorial chromatin state dynamics in space and time. *bioRxiv*, page 38612, jan 2016.
- Trapnell C., Cacchiarelli D., Grimsby J., et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.
- Vaidya V. S., Ramirez V., Ichimura T., Bobadilla N. A., and Bonventre J. V. Urinary kidney injury molecule-1: a sensitive quantitative biomarker for early detection of kidney tubular injury. *American journal of physiology. Renal physiology*, 290(2):F517–29, feb 2006.
- Maaten L.v. d and Hinton G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- Van Dijk D., Sharma R., Nainys J., et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.
- Vavasis S. A. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.

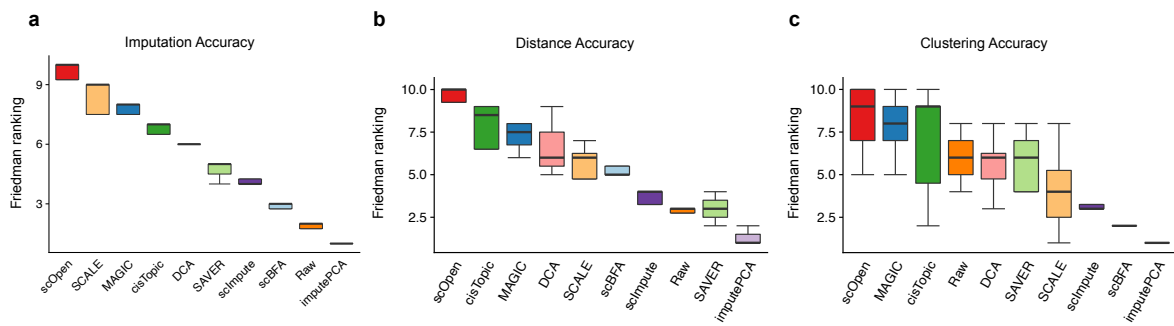


- Virtanen P., Gommers R., Oliphant T. E., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.
- Welch J. D., Kozareva V., Ferreira A., et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, 2019.
- Wolf F. A., Angerer P., and Theis F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.
- Wu H., Kirita Y., Donnelly E. L., and Humphreys B. D. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *Journal of the American Society of Nephrology*, 30(1):23 LP – 32, jan 2019.
- Xie J., Kelley S., and Szymanski B. K. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):1–35, 2013.
- Xiong L., Xu K., Tian K., et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nature Communications*, 10(1):4576, 2019.
- Yoshida H., Lareau C. A., Ramirez R. N., et al. The cis-regulatory atlas of the mouse immune system. *Cell*, 176(4):897–912.e20, 2019.
- Zamanighomi M., Lin Z., Daley T., et al. Unsupervised clustering and epigenetic classification of single cells. *Nature Communications*, 9(1):2410, 2018.
- Zhang K., Hocker J. D., Miller M., et al. A cell atlas of chromatin accessibility across 25 adult human tissues. *bioRxiv*, page 2021.02.17.431699, jan 2021.
- Zhang Y., Liu T., Meyer C. A., et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008.
- Zhang Z., Qin Z., Li P., Yang Q., and Shao J. Multi-view discriminative learning via joint non-negative matrix factorization. In *International Conference on Database Systems for Advanced Applications*, pages 542–557. Springer, 2018.
- Zhu C., Yu M., Huang H., et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural & Molecular Biology*, 26(11):1063–1070, 2019.



## Appendix

### A.1 Technical Validation



**Figure A.1: Overall rank of the imputation methods.** **a**, The Friedman ranking of imputation methods in terms of the average AUPR for each benchmarking dataset. Methods are ordered by median value of ranks. **b**, Same as **a** for distance accuracy. **c**, Same as **a** for clustering accuracy.

## A.1. Technical Validation

Dataset	Cell type	Number of cells	Positives	Negatives
<i>Cell line</i>	BJ	79	23,222	102,425
	GM12878	348	29,954	95,693
	H1ESC	95	16,442	109,205
	HL60	96	9,034	116,613
	K562	510	30,044	95,603
	TF1	96	13,299	112,348
<i>Hematopoiesis</i>	CLP	88	7,726	101,692
	CMP	630	55,791	53,627
	GMP	502	49,760	59,658
	HSC	377	40,217	69,201
	LMPP	93	15,003	94,415
	MEP	188	34,440	74,978
	MPP	162	17,761	91,657
	pDC	170	28,233	81,185
<i>T cells</i>	Jurkat T cell	296	10,573	38,771
	Memory T cell	135	5,341	44,003
	Naive T cell	185	12,645	36,699
	Th17 T cell	149	6,207	43,137
<i>PBMC</i>	CD56 bright NK cells	507	33,378	73,557
	CD56 dim NK cells	472	31,696	75,239
	Classical monocytes	1,929	66,799	40,136
	Effector CD8 T cells	385	35,540	71,395
	Intermediate monocytes	664	57,199	49,736
	MAIT T cells	106	25,548	81,387
	Memory B cells	420	41,046	65,889
	Memory CD4 T cells	1,611	43,931	63,004
	Myeloid DC	242	49,564	57,371
	Naive B cells	295	34,320	72,615
	Naive CD4 T cells	1,462	42,617	64,318
	Naive CD8 T cells	1,549	43,596	63,339
	Non-classical monocytes	383	50,312	56,623
	Plasmacytoid DC	107	34,492	72,443

**Table A.1: Statistics of the positive and negative peaks in each benchmarking dataset.**

	8	7	9	6	10	11	12	5	13	4	14	15	16	17	18	19	20	3	21	22	23	24	25	26	27	28	29	30	2
7																													
9																													
6	**	**																											
10	**	**	**																										
11	**	**	**	**																									
12	**	**	**	**	**																								
5	**	**	**	**	**	**																							
13	**	**	**	**	**	**	**																						
4	**	**	**	**	**	**	**	**																					
14	**	**	**	**	**	**	**	**	**																				
15	**	**	**	**	**	**	**	**	**	**																			
16	**	**	**	**	**	**	**	**	**	**	**		*																
17	**	**	**	**	**	**	**	**	**	**	**	**	**																
18	**	**	**	**	**	**	**	**	**	**	**	**	**	**															
19	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**														
20	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**													
3	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**												
21	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**											
22	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**										
23	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**									
24	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	*								
25	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**							
26	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**						
27	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**						
28	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	*					
29	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**				
30	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**			
2	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**
Raw	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**

**Table A.2: Friedman-Nemenyi test of the number of components for simulation data based on imputation accuracy.** We applied scOpen with different number of components from 2 to 30 and measured the imputation accuracy by AUPR for each single cell. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	10	9	11	15	8	14	13	12	16	17	20	18	24	22	21	23	19	28	25	26	27	30	29	7	5	6	4	3	Raw
9																													
11																													
15																													
8																													
14																													
13																													
12																													
16																													
17																													
20																													
18																													
24																													
22																													
21																													
23																													
19																													
28																													
25	*																												
26	*	*																											
27	**	**	**																										
30	**	**	**																										
29	**	**	**	*																									
7	**	**	**	*																									
5	**	**	**	**	**	**	**	**	**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
6	**	**	**	**	**	**	**	**	**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
4	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**
3	**	**	**	**	**	**	**	**	**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Raw	**	**	**	**	**	**	**	**	**	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
2	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	*

**Table A.3: Friedman-Nemenyi test of the number of components for simulation data based on imputation accuracy.** We applied scOpen with different number of components from 2 to 30 and measured the clustering accuracy by ARI. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	1	2	4	0.1	0.01	0.001	0
2	**						
4	**						
0.1	**	*	*				
0.01	**	**	**	**			
0.001	**	**	**	**			
0	**	**	**	**			
Raw	**	**	**	**	**	**	**

**Table A.4: Friedman-Nemenyi test of the regularization based on imputation accuracy.** We applied scOpen with different number regularization parameters of and measured the imputation accuracy by AUPR. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	0.1	1	0.01	0	0.001	Raw	2
1							
0.01							
0							
0.001							
Raw	*						
2	**	*	*				
4	**	**	**	**	*		

**Table A.5: Friedman-Nemenyi test of the regularization parameters for simulation data based on clustering accuracy.** We applied scOpen with different number regularization parameters of and measured the clustering accuracy by ARI. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	scOpen	SCALE	MAGIC	cisTopic	DCA	scImpute	SAVER	scBFA	Raw
SCALE									
MAGIC	**	**							
cisTopic	**	**							
DCA	**	**	**	**					
scImpute	**	**	**	**					
SAVER	**	**	**	**	**	**			
scBFA	**	**	**	**	**	**	**		
Raw	**	**	**	**	**	**	**	**	
imputePCA	**	**	**	**	**	**	**	**	**

**Table A.6: Friedman-Nemenyi test of the imputation accuracy for Cell line dataset.** We applied the imputation methods on *Cell line* dataset and measured the imputation accuracy by AUPR. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

## A.1. Technical Validation

	<i>scOpen</i>	<i>SCALE</i>	<i>MAGIC</i>	<i>cisTopic</i>	<i>DCA</i>	<i>SAVER</i>	<i>scImpute</i>	<i>scBFA</i>	<i>Raw</i>
<i>SCALE</i>									
<i>MAGIC</i>	**	**							
<i>cisTopic</i>	**	**							
<i>DCA</i>	**	**	**	**					
<i>SAVER</i>	**	**	**	**	**				
<i>scImpute</i>	**	**	**	**	**	**			
<i>scBFA</i>	**	**	**	**	**	**	**		
<i>Raw</i>	**	**	**	**	**	**	**	**	
<i>imputePCA</i>	**	**	**	**	**	**	**	**	**

**Table A.7: Friedman-Nemenyi test of the imputation accuracy for Hematopoiesis dataset.** We applied the imputation methods on *Hematopoiesis* dataset and measured the imputation accuracy by AUPR. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>scOpen</i>	<i>MAGIC</i>	<i>SCALE</i>	<i>cisTopic</i>	<i>DCA</i>	<i>SAVER</i>	<i>scImpute</i>	<i>Raw</i>	<i>scBFA</i>
<i>MAGIC</i>									
<i>SCALE</i>									
<i>cisTopic</i>	**	**	**						
<i>DCA</i>	**	**	**	**					
<i>SAVER</i>	**	**	**	**	**				
<i>scImpute</i>	**	**	**	**	**				
<i>Raw</i>	**	**	**	**	**	**	**		
<i>scBFA</i>	**	**	**	**	**	**	**		
<i>imputePCA</i>	**	**	**	**	**	**	**	**	**

**Table A.8: Friedman-Nemenyi test of the imputation accuracy for T cells dataset.** We applied the imputation methods on *T cells* dataset and measured the imputation accuracy by AUPR. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.



	<i>scOpen</i>	<i>MAGIC</i>	<i>cisTopic</i>	<i>scImpute</i>	<i>SCALE</i>	<i>scBFA</i>
MAGIC	**					
cisTopic	**	**				
scImpute	**	**	**			
SCALE	**	**	**	**		
scBFA	**	**	**	**	**	
Raw	**	**	**	**	**	**

**Table A.9: Friedman-Nemenyi test of the imputation accuracy for PBMC dataset.** We applied the imputation methods on *PBMC* dataset and measured the imputation accuracy by AUPR. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>scOpen</i>	<i>cisTopic</i>	<i>MAGIC</i>	<i>SCALE</i>	<i>DCA</i>	<i>scBFA</i>	<i>scImpute</i>	<i>SAVER</i>	<i>Raw</i>
cisTopic	**								
MAGIC	**								
SCALE	**								
DCA	**	**	**	**					
scBFA	**	**	**	**					
scImpute	**	**	**	**	**	**			
SAVER	**	**	**	**	**	**	**		
Raw	**	**	**	**	**	**	**		
imputePCA	**	**	**	**	**	**	**	**	*

**Table A.10: Friedman-Nemenyi test of the distance accuracy for Cell line dataset.** We applied the imputation methods on *Cell line* dataset and measured the distance accuracy by silhouette score. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

## A.1. Technical Validation

	<i>scOpen</i>	<i>DCA</i>	<i>cisTopic</i>	<i>MAGIC</i>	<i>SAVER</i>	<i>SCALE</i>	<i>imputePCA</i>	<i>Raw</i>	<i>scBFA</i>
DCA	**								
cisTopic	**	**							
MAGIC	**	**	**						
SAVER	**	**	**	**					
SCALE	**	**	**	**					
imputePCA	**	**	**	**					
Raw	**	**	**	**					
scBFA	**	**	**	**	*				
scImpute	**	**	**	**	**	**	**	**	**

**Table A.11: Friedman-Nemenyi test of the distance accuracy for Hematopoiesis dataset.** We applied the imputation methods on *Hematopoiesis* dataset and measured the distance accuracy by silhouette score. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>scOpen</i>	<i>cisTopic</i>	<i>MAGIC</i>	<i>scBFA</i>	<i>SCALE</i>	<i>DCA</i>	<i>scImpute</i>	<i>Raw</i>	<i>SAVER</i>
cisTopic	**								
MAGIC	**								
scBFA	**	**	**						
SCALE	**	**	**						
DCA	**	**	**	**					
scImpute	**	**	**	**	*				
Raw	**	**	**	**	**				
SAVER	**	**	**	**	**	**	**		
imputePCA	**	**	**	**	**	**	**	**	

**Table A.12: Friedman-Nemenyi test of the distance accuracy for T cells dataset.** We applied the imputation methods on *T cells* dataset and measured the distance accuracy by silhouette score. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>scOpen</i>	<i>MAGIC</i>	<i>scBFA</i>	<i>scImpute</i>	<i>cisTopic</i>	<i>Raw</i>
MAGIC	**					
scBFA	**	**				
scImpute	**	**	**			
cisTopic	**	**	**	**		
Raw	**	**	**	**	**	
SCALE	**	**	**	**	**	**

**Table A.13: Friedman-Nemenyi test of the distance accuracy for PBMC dataset.** We applied the imputation methods on *PBMC* dataset and measured the distance accuracy by silhouette score. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>cisTopic</i>	<i>scOpen</i>	<i>MAGIC</i>	<i>DCA</i>	<i>Raw</i>	<i>SCALE</i>	<i>SAVER</i>	<i>scImpute</i>	<i>imputePCA</i>
scOpen									
MAGIC									
DCA									
Raw									
SCALE									
SAVER									
scImpute	**	**	*						
imputePCA	**	**	**	*					
scBFA	**	**	**	*					

**Table A.14: Friedman-Nemenyi test of the clustering accuracy for Cell line dataset.** We applied the imputation methods on *Cell line* dataset and measured the clustering accuracy by ARI. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

## A.1. Technical Validation

	<i>scOpen</i>	<i>cisTopic</i>	<i>MAGIC</i>	<i>Raw</i>	<i>SAVER</i>	<i>DCA</i>	<i>SCALE</i>	<i>scImpute</i>	<i>scBFA</i>
<i>cisTopic</i>									
<i>MAGIC</i>									
<i>Raw</i>									
<i>SAVER</i>									
<i>DCA</i>									
<i>SCALE</i>	*	*							
<i>scImpute</i>	**	**	*						
<i>scBFA</i>	**	**	**						
<i>imputePCA</i>	**	**	**	*	*	*			

**Table A.15: Friedman-Nemenyi test of the clustering accuracy for Hematopoiesis dataset.** We applied the imputation methods on *Hematopoiesis* dataset and measured the clustering accuracy by ARI. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>scOpen</i>	<i>MAGIC</i>	<i>cisTopic</i>	<i>SAVER</i>	<i>Raw</i>	<i>SCALE</i>	<i>DCA</i>	<i>scImpute</i>	<i>scBFA</i>
<i>MAGIC</i>									
<i>cisTopic</i>									
<i>SAVER</i>									
<i>Raw</i>									
<i>SCALE</i>									
<i>DCA</i>	*	*							
<i>scImpute</i>	**	**	*						
<i>scBFA</i>	**	**	**						
<i>imputePCA</i>	**	**	**	*	*				

**Table A.16: Friedman-Nemenyi test of the clustering accuracy for T cells dataset.** We applied the imputation methods on *T cells* dataset and measured the clustering accuracy by ARI. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>scOpen</i>	<i>Raw</i>	<i>MAGIC</i>	<i>scImpute</i>	<i>scBFA</i>	<i>cisTopic</i>
Raw						
MAGIC						
scImpute						
scBFA	**	*				
cisTopic	**	*				
SCALE	**	**	**			

**Table A.17: Friedman-Nemenyi test of the clustering accuracy for PBMC dataset.** We applied the imputation methods on *PBMC* dataset and measured the clustering accuracy by ARI. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>cisTopic</i>	<i>scOpen-impute</i>	<i>scOpen</i>	<i>cisTopic-impute</i>	<i>SnapATAC</i>
scOpen-impute	**				
scOpen	**	**			
cisTopic-impute	**	**	**		
SnapATAC	**	**	**	**	
Cusanovich2018	**	**	**	**	

**Table A.18: Friedman-Nemenyi test of the distance accuracy of dimensionality reduction methods for Cell line dataset.** We applied the dimensionality reduction methods on *Cell line* dataset and measured the distance accuracy by silhouette score. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

	<i>scOpen-impute</i>	<i>scOpen</i>	<i>cisTopic</i>	<i>cisTopic-impute</i>	<i>SnapATAC</i>
scOpen	**				
cisTopic	**	**			
cisTopic-impute	**	**	**		
SnapATAC	**	**	**		
Cusanovich2018	**	**	**	**	**

**Table A.19: Friedman-Nemenyi test of the distance accuracy of dimensionality reduction methods for Hematopoiesis dataset.** We applied the dimensionality reduction methods on *Hematopoiesis* dataset and measured the distance accuracy by silhouette score. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

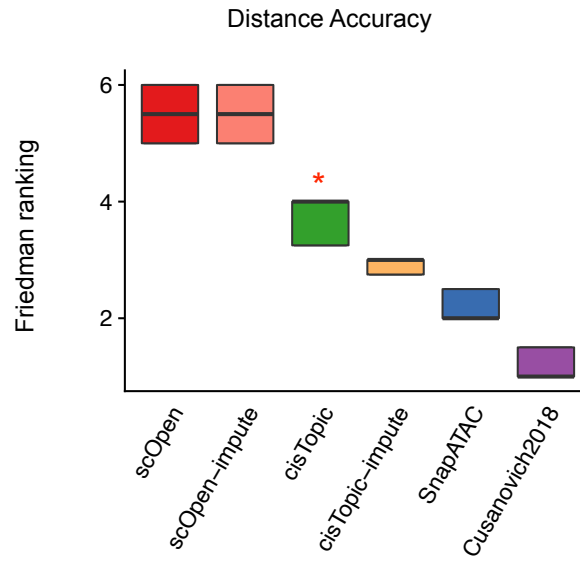
## A.1. Technical Validation

	<i>scOpen</i>	<i>cisTopic</i>	<i>scOpen-impute</i>	<i>cisTopic-impute</i>	<i>SnapATAC</i>
<i>cisTopic</i>	**				
<i>scOpen-impute</i>	**				
<i>cisTopic-impute</i>	**	**	**		
<i>SnapATAC</i>	**	**	**	**	
<i>Cusanovich2018</i>	**	**	**	**	**

**Table A.20: Friedman-Nemenyi test of the distance accuracy of dimensionality reduction methods for T cells dataset.** We applied the dimensionality reduction methods on *T cells* dataset and measured the distance accuracy by silhouette score. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.

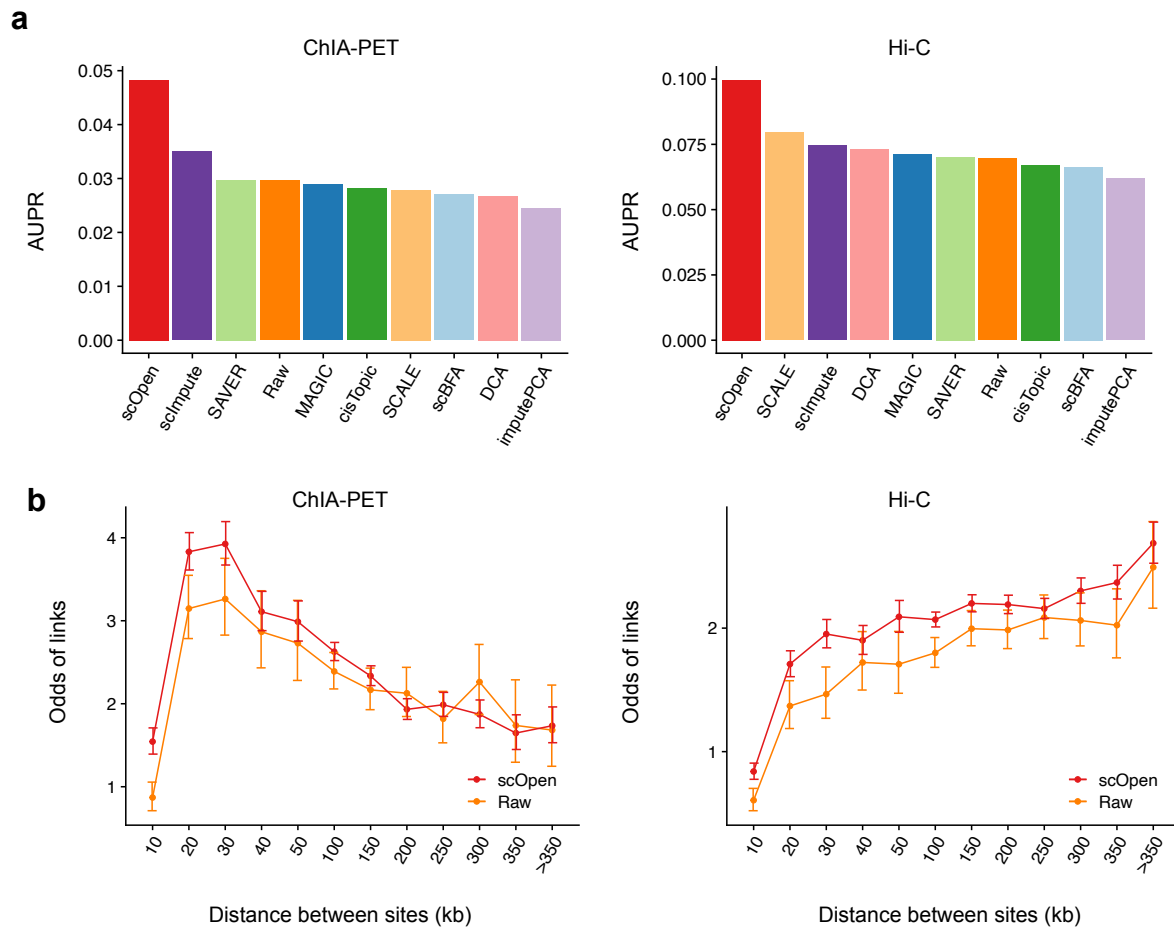
	<i>scOpen</i>	<i>scOpen-impute</i>	<i>Cusanovich2018</i>	<i>SnapATAC</i>	<i>cisTopic-impute</i>
<i>scOpen-impute</i>	**				
<i>Cusanovich2018</i>	**	**			
<i>SnapATAC</i>	**	**	**		
<i>cisTopic-impute</i>	**	**	**	**	
<i>cisTopic</i>	**	**	**	**	**

**Table A.21: Friedman-Nemenyi test of the distance accuracy of dimensionality reduction methods for PBMC dataset.** We applied the dimensionality reduction methods on *PBMC* dataset and measured the distance accuracy by silhouette score. The asterisk and the two asterisks, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.01.



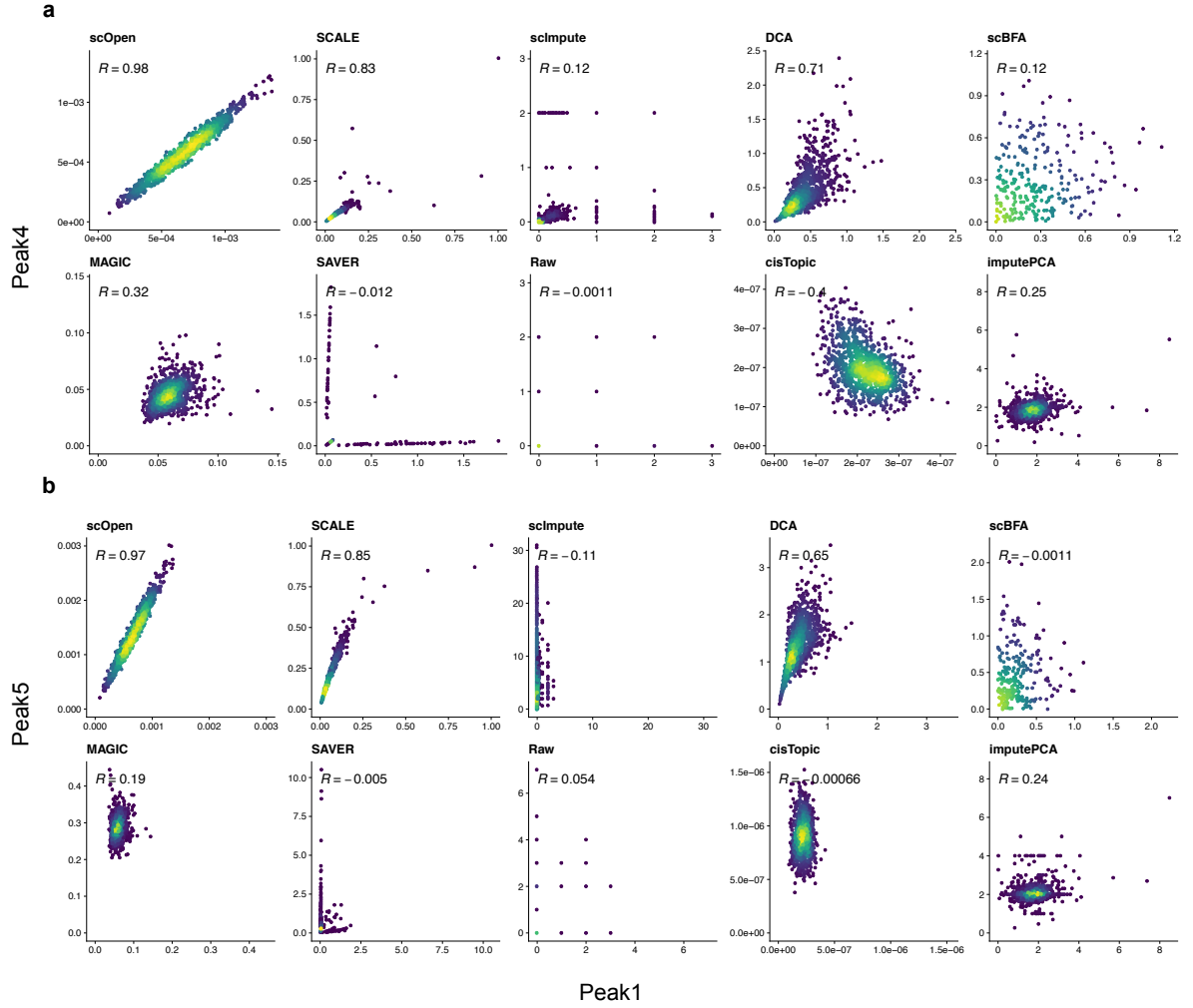
**Figure A.2: The combined rank of the dimensionality reduction methods based on distance and clustering accuracy.** The Friedman ranking of dimensionality reduction methods in terms of the average silhouette score for each benchmarking dataset. Methods are ordered by median value of ranks. Wilcoxon Rank Sum test was used to compare scOpen with scOpen-impute and MAGIC. The asterisk means that the method is outperformed by scOpen with significance level of 0.05.

## A.1. Technical Validation



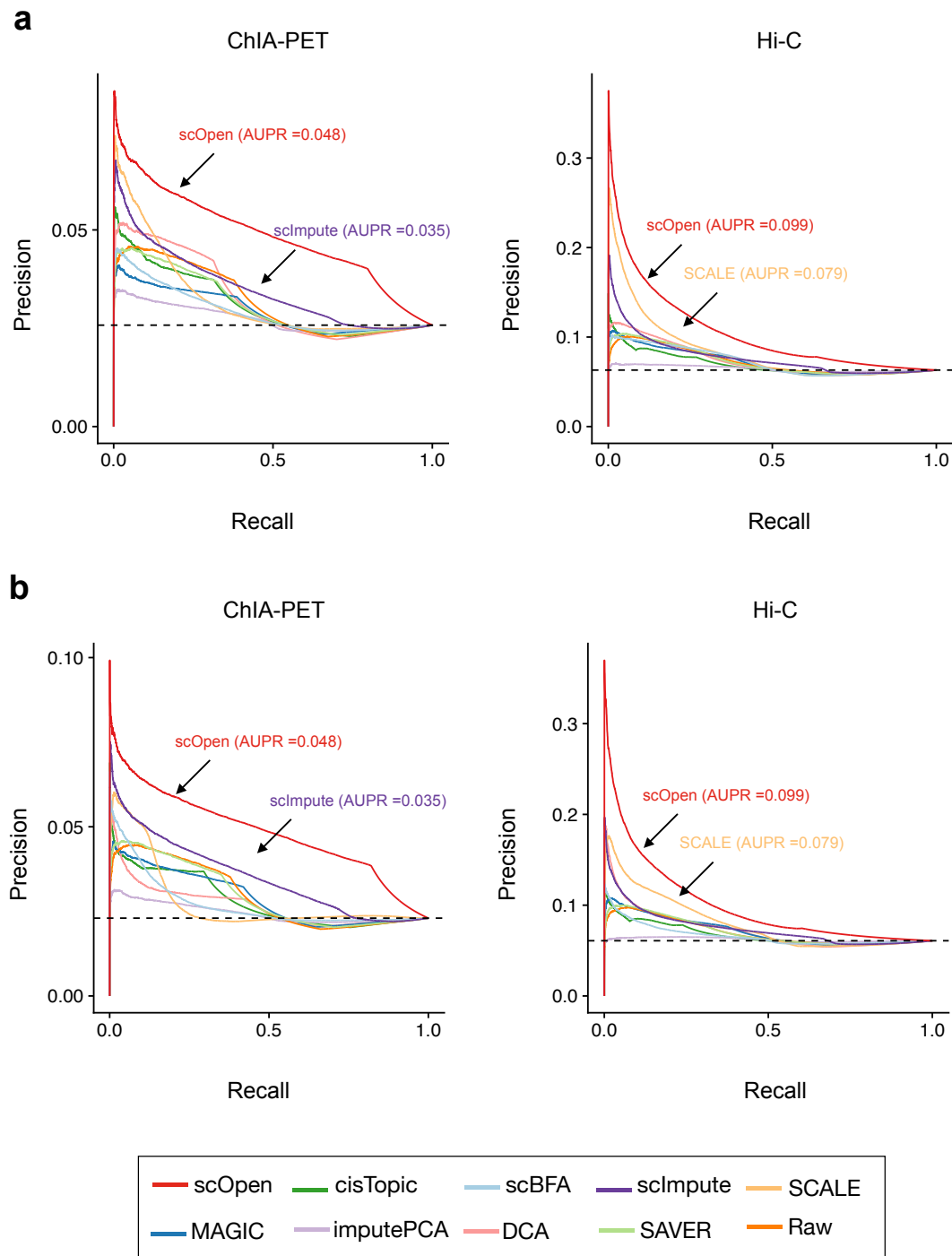
**Figure A.3: Evaluation of the predicted interactions by Cicero.** **a**, Barplots showing AUPR of the predicted peak-to-peak co-accessibility links using either raw or imputed matrix with GM12878 single-cell ATAC-seq data. Analysis was executed with Cicero. Left, links are evaluated using ChIA-PET data as true labels. Right, links are evaluated using Hi-C data as true labels. **b**, Odds ratio (y-axis) of Cicero predicted co-accessible sites ( $n = 3,853,260$ ) also supported by pol-II ChIA-PET (left) and Hi-C (right) vs. distance between sites (x-axis). Error bars indicate 95% confidence intervals calculated using Fisher's exact test. Odds ratio superior than 1 indicates a positive relationship.





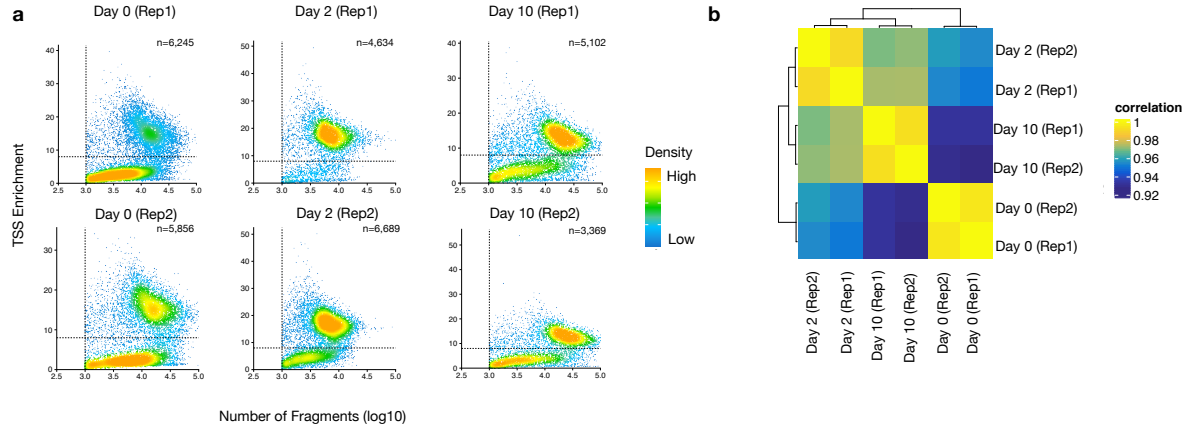
**Figure A.4: Visualization of co-accessibility score a**, Scatter plot showing single-cell accessibility scores estimated by top-performing imputation methods for the link between peak 1 and peak 4. Each dot represents a cell and color refers to density. Pearson correlation is shown on the left-upper corner. **b**, Same as **a** for peak 1 and peak5.

## A.1. Technical Validation



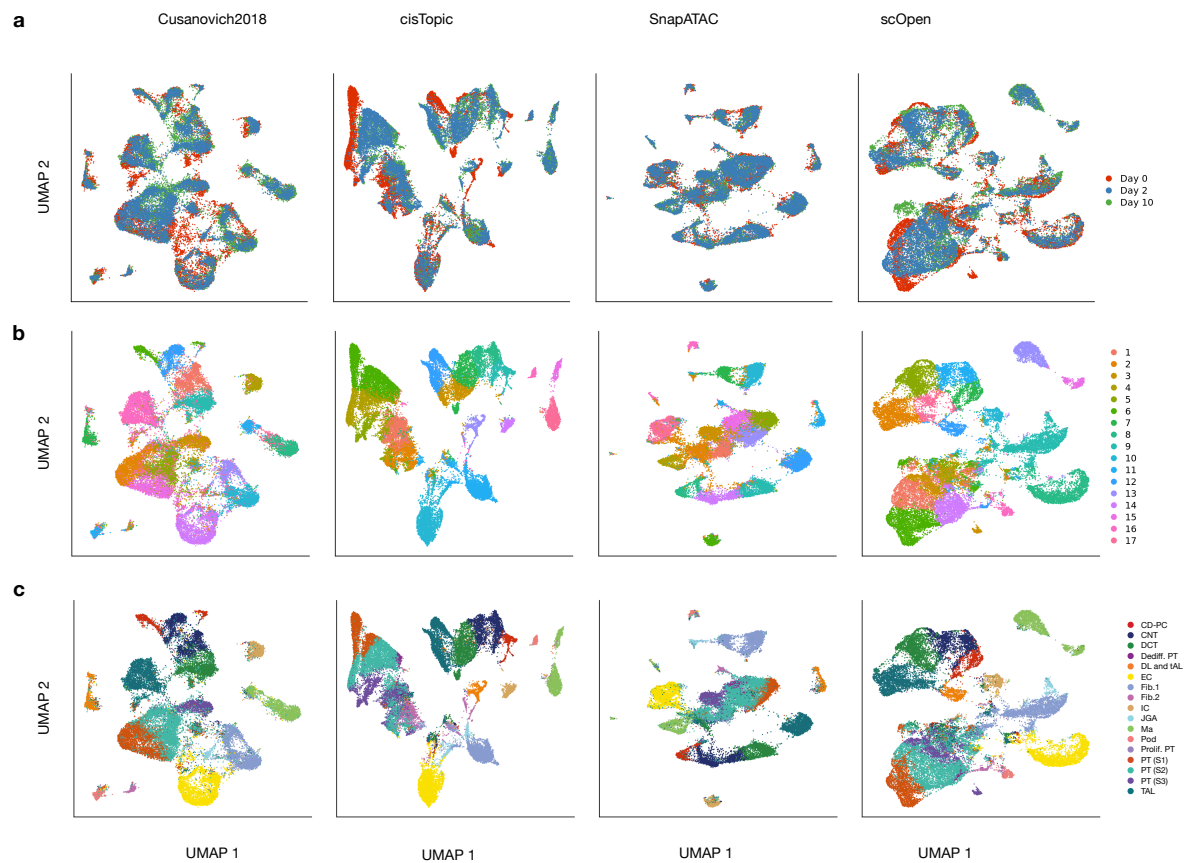
**Figure A.5: Evaluation of the predicted interactions by Cicero using down-sampled data. a,** Precision-recall curves showing the evaluation of the predicted links on GM12878 cells using the raw and imputed matrix as input after down-sampling to 50%. **b** Same as **a** after down-sampling to 25%.

## A.2 Biological Validation

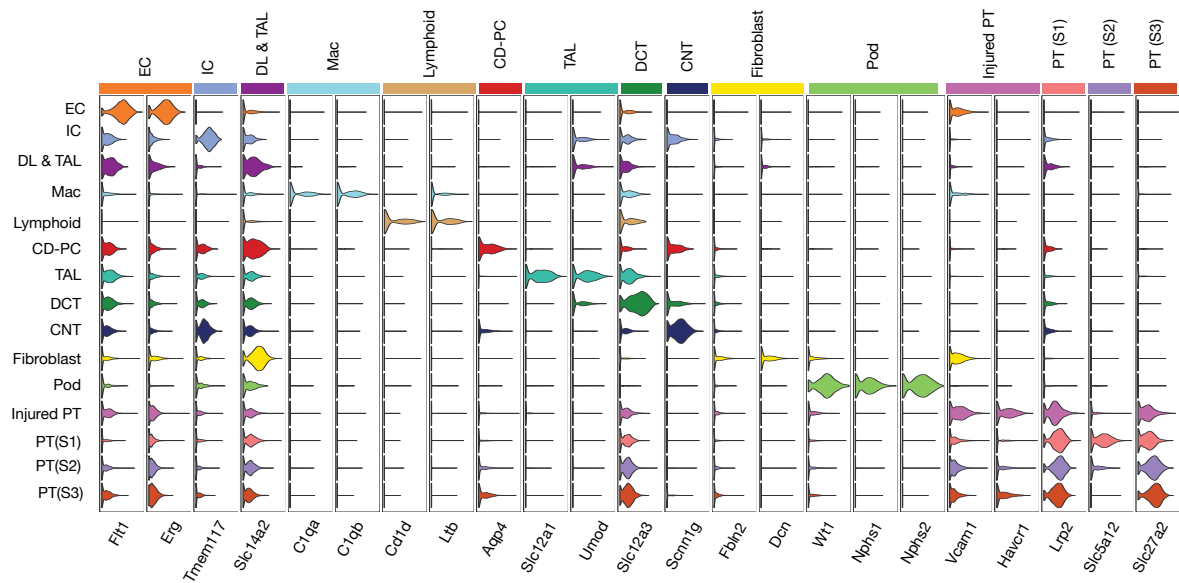


**Figure A.6: Quality control of UO scATAC-seq data.** **a**, Scatter plot showing number of unique fragments vs. TSS enrichment of UO scATAC-seq for each sample. Each dot represents a cell and the dash lines represent cut-off used for cell filtering. The number of cells that pass filtering is shown on right-upper corner. **b**, Heatmap showing the correlation between samples.

## A.2. Biological Validation

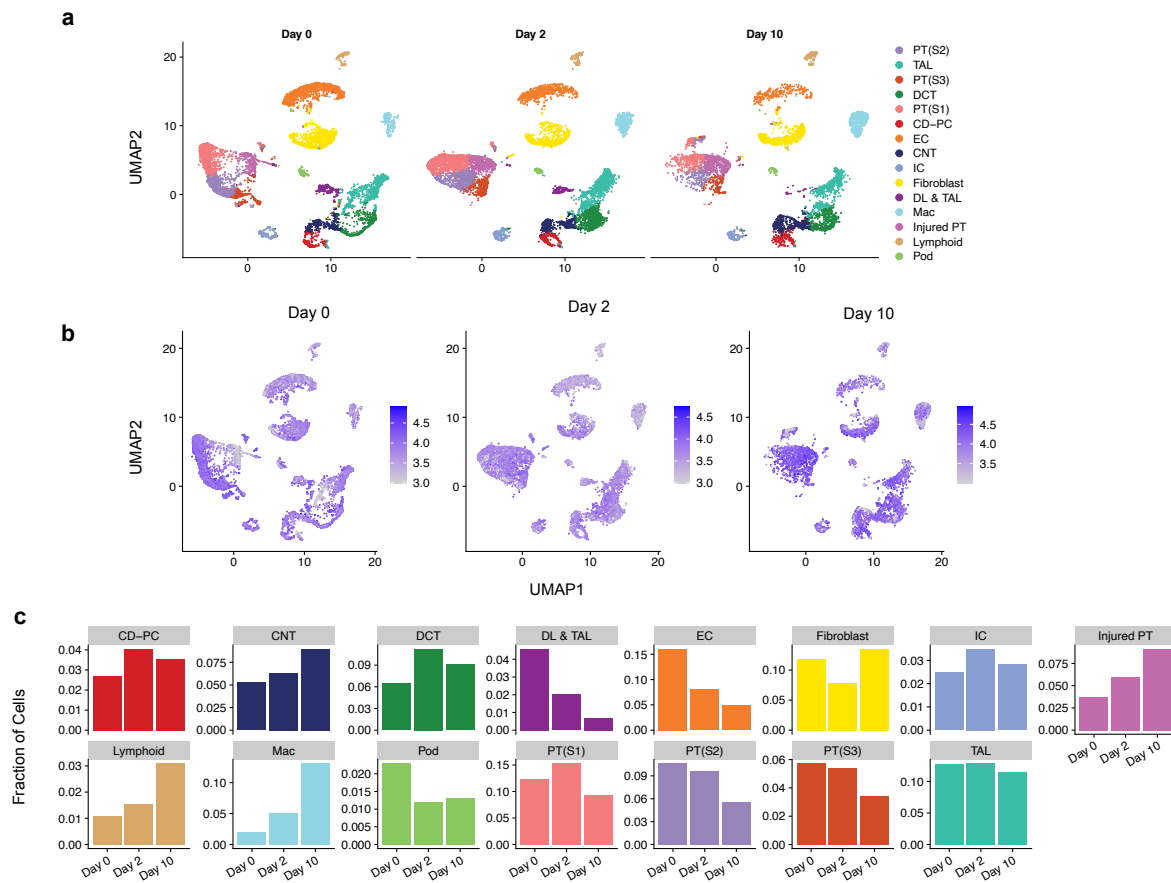


**Figure A.7: Visualization of integrated data by using different dimensionality reduction methods.** **a**, UMAP plots showing data integration results by using different scATAC-seq dimension reduction/clustering pipelines. Each dot represent a cell and cells are colored by different time points. **b**, UMAP plots showing clustering results for each dimension reduction method. Cells are colored by clusters. **c**, UMAP plots showing label transferred results from a public UUO scRNA-seq dataset. Cells are colored by predicted labels.

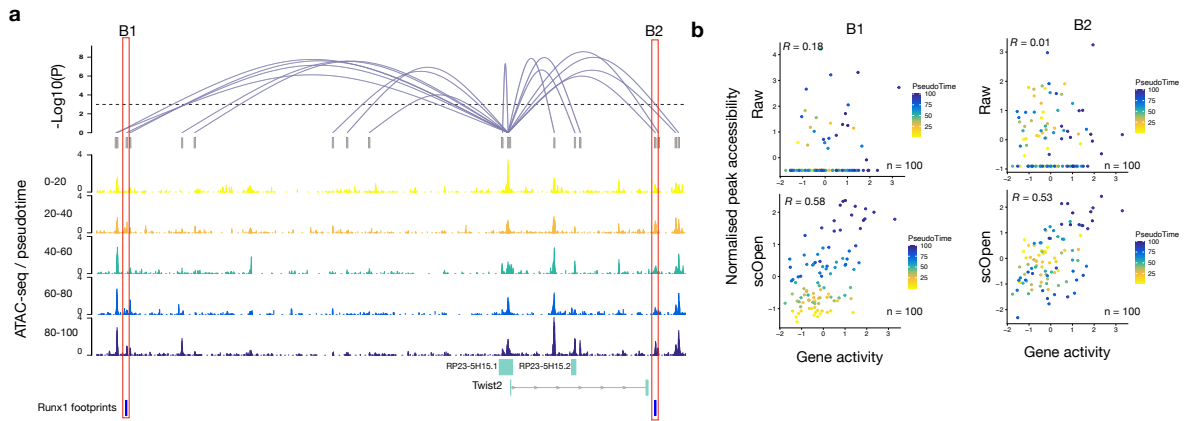


**Figure A.8: Visualization of marker genes for each annotated cell type.** Violin plot showing cluster specific (y-axis) gene accessibility score associated to known marker genes for kidney cells (x-axis).

## A.2. Biological Validation



**Figure A.9: Visualization of annotated cell types for each time point.** **a**, Scatter plots showing condition-specific UMAP visualization of UO scATAC-seq data for each sample. **b**, Visualization of the data quality for each sample. Colors refer to number of fragments per cell after log10 transformation. **c**, Bar plots showing proportion of each cell type across different time points.



**Figure A.10: Twist2 is regulated by Runx1 in myofibroblast.** **a** Peak-to-Gene links (top) predicted on scOpen matrix and associated to Twist2 in fibroblast cells. The height of links represents its significance. Dash line represents the threshold of significance (FDR = 0.001). ATAC-seq tracks (below) were generated from pseudo-bulk profiles of fibroblast/myofibroblast cells with increasing pseudo time (0-20, 20-40, 40-60, 60-80, and 80-100). Binding sites of Runx1 (B1-B2) supported by ATAC-seq footprints and overlapping to peaks are highlighted on the bottom. **b** Scatter plot showing gene activity of Twist2 and normalized peak accessibility from raw (upper) or scOpen imputed matrix (lower) for peak-to-gene link B1 and B2. Each dot represents cells in a given pseudotime and the overall correlation is shown in the left-upper corner.