MACHINE LEARNING AND MATHEMATICAL MODELS FOR SINGLE-CELL DATA ANALYSIS

EDITED BY: Le Ou-Yang, Xiaofei Zhang, Jiajun Zhang, Jin Chen and Min Wu
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-83250-184-9
DOI 10.3389/978-2-83250-184-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

MACHINE LEARNING AND MATHEMATICAL MODELS FOR SINGLE-CELL DATA ANALYSIS

Topic Editors:
Le Ou-Yang, Shenzhen University, China
Xiaofei Zhang, Central China Normal University, China
Jiajun Zhang, Sun Yat-sen University, China
Jin Chen, University of Kentucky, United States
Min Wu, Institute for Infocomm Research (A*STAR), Singapore

Citation: Ou-Yang, L., Zhang, X., Zhang, J., Chen, J., Wu, M., eds. (2022). Machine Learning and Mathematical Models for Single-Cell Data Analysis. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-184-9

Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 03 June 2022
doi: 10.3389/fgene.2022.911999

Editorial: Machine Learning and Mathematical Models for Single-Cell Data Analysis

Le Ou-Yang[1], Xiao-Fei Zhang[2], Jiajun Zhang[3], Jin Chen[4] and Min Wu[5]*

[1]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Guangdong Laboratory of Artificial Intelligence and Digital Economy(SZ), College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China, [2]School of Mathematics and Statistics and Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan, China, [3]Guangdong Province Key Laboratory of Computational Science, School of Mathematics, Sun Yat-sen University, Guangzhou, China, [4]Institute for Biomedical Informatics, University of Kentucky, Lexington, KY, United States, [5]Institute for Infocomm Research (I2R), A*STAR, Singapore, Singapore

Keywords: single-cell omics data, machine learning, mathematical modelling, data integration, network modeling

Editorial on the Research Topic

Machine Learning and Mathematical Models for Single-Cell Data Analysis

Understanding how individual cells communicate with each other and respond to evolution and perturbations is a central challenge of biology (Altschuler and Wu, 2010). Due to the heterogeneity of cells, studying a bulk population of cells may confound the variability of cell-type compositions, single cell analysis has the potential to enable a more systematic study of the inner workings of biological systems, and allows us to uncover the underlying mechanisms for cellular functions and biological processes such as cell differentiation and disease development. In the past decade, advances in single-cell isolation and sequencing technologies have enabled the assay of DNA, mRNA, and protein abundances at single-cell resolution, which promote the study of genomics, transcriptomics, proteomics and metabolomics at the sinlge cell level. For example, single-cell genomic analysis can shed light to the genomic variability of individual cells, while single-cell transcriptomic and proteomic analysis can help to reveal the types and functional states of individual cells (Shapiro et al., 2013). However, processing single-cell data of high dimensionality and scale is inherently difficult, especially considering the degree of noise, sparsity, batch effects and heterogeneity in the data (Amodio et al., 2019). Thus, there is an urgent need for developing computational models which can handle the size, dimensionality, and various characteristics of single-cell data. In this Research Topic of Frontiers in Genetics on "*Machine Learning and Mathematical Models for Single-Cell Data Analysis*," we have collected eight manuscripts that used machine learning algorithms or mathematical models to solve problems in single cell analysis.

Single-cell and whole tissue RNA sequencing technologies enable the Research Topic of detailed information about biological processes at genomic and transcriptomic levels. Besides, existing microscopy and cell-resolution imaging techniques allow the high-quality characterization of morphology and physiology at the level of extended fragments of tissues and organs. Bobrovskikh et al. summarized the potential of single-cell technologies together with advanced imaging techniques for computational modelling in plants. They reviewed currently available single-cell data analysis approaches, advanced imaging technologies in plant research with single-cell resolution and cell-based modelling approaches. They shown how the combination of single-cell data, morphometric data and cell-based models help to expand the understanding of tissue and organ morphogenesis.

Tissues are constituted of heterogeneous cell types. Although single-cell RNA sequencing has paved the way to a deeper understanding of organismal cellular composition, the high cost and technical noise have prevented its wide application. As an alternative, computational deconvolution

Edited and reviewed by:
Alfredo Pulvirenti,
University of Catania, Italy

***Correspondence:**
Le Ou-Yang
leouyang@szu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *03 April 2022*
Accepted: *19 May 2022*
Published: *03 June 2022*

Citation:
Ou-Yang L, Zhang X-F, Zhang J, Chen J and Wu M (2022) Editorial: Machine Learning and Mathematical Models for Single-Cell Data Analysis. Front. Genet. 13:911999. doi: 10.3389/fgene.2022.911999

of bulk tissues can be a cost-effective solution (Jin and Liu, 2021). Liu et al. proposed a deconvolution method, named DecOT, to characterize the cell type composition from bulk tissue RNA-seq data. DecOT uses the optimal transport distance as a loss and applies an ensemble framework to integrate reference information from scRNA-seq data of multiple individuals. Experiment results on real data sets demonstrated that DecOT outperformed other existing methods and was robust to the choice of references.

The development of single-cell sequencing technologies promotes the researches on developmental physiology and disease (Potter, 2018), but the spatial information of individual cells is lost due to the tissue dissociation processes in these technologies. Highly multiplexed imaging technologies, such as imaging mass cytometry (IMC), are powerful tools to exploit the composition and interactions of cells in tumor microenvironments at subcellular resolution. However, due to the high resolution and large number of channels, how to process and interpret IMC image data still remains challenging (Chang et al., 2017). To improve the accuracy of single cell segmentation, which is a critical step to process IMC image data, Xiao et al. developed a deep neural network (DNN)-based cell segmentation method, named Dice-XMBD. Dice-XMBD is marker agnostic and can perform accurate cell segmentation of IMC images of different channel configurations without modification.

Advances in single-cell RNA-sequencing (scRNA-seq) technology provided an unprecedented opportunity for researchers to study the identity and mechanisms of single cells (Morris, 2019). Besides scRNA-seq data, spatial location data can also provide important information on the cells' micro-environment and cell-cell interactions (Mayr et al., 2019), which can contribute to cell type identification. Oh et al. proposed a hybrid clustering approach, named single-cell Hybrid Nonnegative Matrix Factorization (scHybridNMF), to perform cell clustering by jointly processing cell location and gene expression data. ScHybridNMF combines sparse nonnegative matrix factorization (sparse NMF) with k-means clustering to cluster high-dimensional gene expression and low-dimensional location data. Experiment results on simulated and real data sets demonstrate the effectiveness of scHybridNMF in detecting cell clusters.

The communication between cells plays a vital role in the development, physicology, and pathology of muticellular organisms. Single-cell RNA-sequencing (scRNA-seq), which measures the expression levels of a great number of genes across various cell types at single-cell resolution, provides a great opportunity to study the cell-cell communication between interacting cells and the signaling response governed by intracellular gene regulatory networks (GRNs) (Shao et al., 2020). Identification the changes of intercellular signaling across different conditions is crucial for understanding how distinct cell states respond to evolution, perturbations, and diseases. Wang et al. generalized their previously developed tool CellChat to enable a flexible comparison analysis of cell-cell communication networks across multiple conditions, which facilitated the detection of signaling changes of cell-cell communication in response to biological perturbations. By studying the signaling changes across three mouse embryonic developmental stages, four time points after mouse spinal cord injury, and patients with different COVID-19 severities (i.e., control, moderate, and critical cases), they verified the effectiveness of their proposed approaches. To infer the changes of GRNs between two different states, Liu et al. proposed a general differential network inference framework, named weighted joint sparse penalized D-trace model (WJSDM). WJSDM can directly infer the differential network between two different states by integrating multi-platform gene expression data and various existing biological knowledge. By applying WJSDM to the gene expression data of ovarian cancer and the scRNA-seq data of circulating tumor cells of prostate cancer, and infer the differential network associated with platinum resistance of ovarian cancer and anti-androgen resistance of prostate cancer, the authors found some important biological insights about the mechanisms underlying platinum resistance of ovarian cancer and anti-androgen resistance of prostate cancer.

Recent advances in experimental biology have generated huge amounts of data. For example, Microwell-Seq, a single-cell RNA-sequencing technology, has been used to analyze the transcriptome of more than 400,000 mouse single cells, covering all major mouse organs (Han et al., 2018). There is an urgent need for next generation methods to deal with large, heterogeneous and complex data sets Camacho et al. (2018). As a promising data processing method, deep learning methods have been employed in biological data processing (Eraslan et al., 2019). However, the deep learning methods usually run as a "black box," which is hard to interpret. The capsule network (CapsNet) is a newly developed deep learning model for digital recognition tasks (Sabour et al., 2017). Wang et al. (2020) proposed a modified CapsNet model, called single cell capsule network (scCapsNet), which is a highly interpretable cell type classifier, with the capability of revealing cell type associated genes by model internal parameters. Based on CapsNet and scCapsNet, Wang et al. proposed a deep learning classifier and data integrator, named MultiCapsNet. The MultiCapsNet model could integrate multiple input sources and standardize the inputs, then use the standardized information for classification through capsule network. The experiment results on three data sets with different data type and application scenarios proved the validity and interpretability of MultiCapsNet.

Cancer immunotherapy has shown to elicit substantial response to many cancers and has led to significant increases in quality of life for cancer patients. This is especially true of checkpoint therapy, which causes tumor regression in previously untreatable cancers. However, the potential mechanisms of checkpoint therapy are still being investigated and there are as of yet few prognostic markers for response (Bai et al., 2020). Immune checkpoint therapies such as PD-1 blockade have vastly improved the treatment of numerous cancers, including basal cell carcinoma (BCC). However, patients afflicted with pancreatic ductal carcinoma (PDAC), one of the deadliest malignancies, overwhelmingly exhibit negative responses to checkpoint therapy. Liu et al. sought to combine data analysis and machine learning to differentiate the putative mechanisms of BCC and PDAC non-response. By comparing two recent single-

cell transcriptomic datasets of PDAC and BCC, the authors identified some potential biomarkers and mechanisms related to BCC and PDAC non-response. By utilizing machine learning classification algorithms, they also discovered that PDAC displays greater similarities to melanoma, which is highly immunogenic and undergoes rapid metastasis, than to BCC (Dollinger et al., 2020).

In summary, this Research Topic covers various aspects of machine learning models, including supervised and unsupervised approaches and their applications for single-cell data analysis, which paves the way for using machine learning and mathematical models in service of various tasks towards single cell analysis. We hope the readers from bioinformatics and the domain specific researchers will be benefitted by reading articles included in this Research Topic.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Altschuler, S. J., and Wu, L. F. (2010). Cellular Heterogeneity: Do Differences Make a Difference? *Cell.* 141, 559–563. doi:10.1016/j.cell.2010.04.033

Amodio, M., Van Dijk, D., Srinivasan, K., Chen, W. S., Mohsen, H., Moon, K. R., et al. (2019). Exploring Single-Cell Data with Deep Multitasking Neural Networks. *Nat. Methods* 16, 1139–1145. doi:10.1038/s41592-019-0576-7

Bai, R., Lv, Z., Xu, D., and Cui, J. (2020). Predictive Biomarkers for Cancer Immunotherapy with Immune Checkpoint Inhibitors. *Biomark. Res.* 8, 34–17. doi:10.1186/s40364-020-00209-0

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation Machine Learning for Biological Networks. *Cell.* 173, 1581–1592. doi:10.1016/j.cell.2018.05.015

Chang, Q., Ornatsky, O. I., Siddiqui, I., Loboda, A., Baranov, V. I., and Hedley, D. W. (2017). Imaging Mass Cytometry. *Cytometry* 91, 160–169. doi:10.1002/cyto.a.23053

Dollinger, E., Bergman, D., Zhou, P., Atwood, S. X., and Nie, Q. (2020). Divergent Resistance Mechanisms to Immunotherapy Explain Responses in Different Skin Cancers. *Cancers* 12, 2946. doi:10.3390/cancers12102946

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep Learning: New Computational Modelling Techniques for Genomics. *Nat. Rev. Genet.* 20, 389–403. doi:10.1038/s41576-019-0122-6

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell.* 172, 1091–1107. doi:10.1016/j.cell.2018.02.001

Jin, H., and Liu, Z. (2021). A Benchmark for Rna-Seq Deconvolution Analysis under Dynamic Testing Environments. *Genome Biol.* 22, 102–123. doi:10.1186/s13059-021-02290-6

Mayr, U., Serra, D., and Liberali, P. (2019). Exploring Single Cells in Space and Time during Tissue Development, Homeostasis and Regeneration. *Development* 146, dev176727. doi:10.1242/dev.176727

Morris, S. A. (2019). The Evolving Concept of Cell Identity in the Single Cell Era. *Development* 146, dev169748. doi:10.1242/dev.169748

Potter, S. S. (2018). Single-cell Rna Sequencing for the Study of Development, Physiology and Disease. *Nat. Rev. Nephrol.* 14, 479–492. doi:10.1038/s41581-018-0021-7

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic Routing between Capsules. *Adv. neural Inf. Process. Syst.* 30, 344. doi:10.1097/01.asw.0000521116.18779.7c

Shao, X., Lu, X., Liao, J., Chen, H., and Fan, X. (2020). New Avenues for Systematically Inferring Cell-Cell Communication: through Single-Cell Transcriptomics Data. *Protein Cell.* 11, 866–880. doi:10.1007/s13238-020-00727-5

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science. *Nat. Rev. Genet.* 14, 618–630. doi:10.1038/nrg3542

Wang, L., Nie, R., Yu, Z., Xin, R., Zheng, C., Zhang, Z., et al. (2020). An Interpretable Deep-Learning Architecture of Capsule Networks for Identifying Cell-type Gene Expression Programs from Single-Cell Rna-Sequencing Data. *Nat. Mach. Intell.* 2, 693–703. doi:10.1038/s42256-020-00244-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 21 May 2021
doi: 10.3389/fgene.2021.652974

A Sight on Single-Cell Transcriptomics in Plants Through the Prism of Cell-Based Computational Modeling Approaches: Benefits and Challenges for Data Analysis

Aleksandr Bobrovskikh[1,2†], *Alexey Doroshkov*[1,3†], *Stefano Mazzoleni*[2], *Fabrizio Cartenì*[2], *Francesco Giannino*[2] *and Ulyana Zubairova*[1,3*]

[1] Laboratory of Plant Growth Biomechanics, Institute of Cytology and Genetics Siberian Branch of Russian Academy of Sciences (SB RAS), Novosibirsk, Russia, [2] Department of Agricultural Sciences, University of Naples Federico II, Naples, Italy, [3] Department of Natural Sciences, Novosibirsk State University, Novosibirsk, Russia

OPEN ACCESS

Edited by:
Le Ou-Yang,
Shenzhen University, China

Reviewed by:
Samuel Seaver,
Argonne National Laboratory (DOE), United States
Markus M. Hilscher,
Science for Life Laboratory (SciLifeLab), Sweden

***Correspondence:**
Ulyana Zubairova
ulyanochka@bionet.nsc.ru

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *13 January 2021*
Accepted: *20 April 2021*
Published: *21 May 2021*

Citation:
Bobrovskikh A, Doroshkov A, Mazzoleni S, Cartenì F, Giannino F and Zubairova U (2021) A Sight on Single-Cell Transcriptomics in Plants Through the Prism of Cell-Based Computational Modeling Approaches: Benefits and Challenges for Data Analysis. Front. Genet. 12:652974. doi: 10.3389/fgene.2021.652974

Single-cell technology is a relatively new and promising way to obtain high-resolution transcriptomic data mostly used for animals during the last decade. However, several scientific groups developed and applied the protocols for some plant tissues. Together with deeply-developed cell-resolution imaging techniques, this achievement opens up new horizons for studying the complex mechanisms of plant tissue architecture formation. While the opportunities for integrating data from transcriptomic to morphogenetic levels in a unified system still present several difficulties, plant tissues have some additional peculiarities. One of the plants' features is that cell-to-cell communication topology through plasmodesmata forms during tissue growth and morphogenesis and results in mutual regulation of expression between neighboring cells affecting internal processes and cell domain development. Undoubtedly, we must take this fact into account when analyzing single-cell transcriptomic data. Cell-based computational modeling approaches successfully used in plant morphogenesis studies promise to be an efficient way to summarize such novel multiscale data. The inverse problem's solutions for these models computed on the real tissue templates can shed light on the restoration of individual cells' spatial localization in the initial plant organ—one of the most ambiguous and challenging stages in single-cell transcriptomic data analysis. This review summarizes new opportunities for advanced plant morphogenesis models, which become possible thanks to single-cell transcriptome data. Besides, we show the prospects of microscopy and cell-resolution imaging techniques to solve several spatial problems in single-cell transcriptomic data analysis and enhance the hybrid modeling framework opportunities.

Keywords: single-cell transcriptomics, cell-based computational models, plant morphogenesis, hybrid modeling approach, modeling software, bioimaging, spatial gene expression maps, systems biology

1. INTRODUCTION

Modern biology is going through the era of big data and omics technologies. Single-cell sequencing (SCS) is one of the breakthroughs and rapidly developing technologies. This technology's value is difficult to overestimate since it allows one to describe with high accuracy the trajectories of cell development and characterize individual cell types (Trapnell, 2015). A targeted study of isolated cells is of particular importance in the context of systems biology, as demonstrated on root hair cells (Hossain et al., 2015). The main steps of SC analysis include cellular dissociation, single-cell RNA sequencing (scRNA-seq), dimensionality reduction, clustering, and reconstruction of the developmental trajectories. McFaline-Figueroa et al. (2020) provide currently available techniques for such kind of analysis. However, such a data-driven approach provides only a partial understanding of the developmental processes for different cell types since it includes only the molecular level.

Thus, a combination of microscopy methods (Li et al., 2014) and imaging techniques (Omari et al., 2020) could provide a new level of understanding the developmental processes. In turn, the combination of high-precision SCS approaches with high-quality microscopic data can be integrated into mathematical models describing morphogenesis. Therefore, we believe that current methods for processing SC data should be coupled with morphological data on a tissue level and computational frameworks describing tissue development. Such a systemic-biological cycle will allow researchers to find out the essential spatiotemporal regulators of morphogenetic processes and provide an *in silico - in vivo* verification of emerging hypotheses.

The relationship between growth characteristics of individual cells and organogenesis was noted in the work of Hong et al. (2018). In particular, it was shown that growth rate and growth direction significantly affect organ developmental processes, and, therefore, could determine the invariant organ formations. Consequently, it is essential to study cells' individual characteristics to create a holistic picture of morphogenetic processes at the tissue and organ levels. The main drivers of morphogenesis are shown schematically below, in **Figure 1**. Stem cells can divide, either symmetrically or with precise daughter-cell size ratio, the so-called formative divisions, which are fundamental determinants in the processes of morphogenesis Smolarkiewicz and Dhonukshe (2013). Also, the emergence of cellular patterns forming tissues significantly depends on the anisotropic cell growth biomechanics, which occurs, in particular, in tip-growing cells (Rounds and Bezanilla, 2013).

In addition to the mechanical factors influencing growth, it is known that the formation of apical meristems (which are the niches of undifferentiated stem cells) is complex and includes molecular, hormonal and epigenetic levels of regulation (Ali et al., 2020). Moreover, the realization of the cell death program is known to be a stimulating factor for hormone signaling in developmental processes (Xuan et al., 2016), and a detailed overview and classification of plant cell death can be found in Locato and De Gara (2018).

The multilevel nature of morphogenetic processes increases the need for systemic biological research that integrates multilevel data. For example, a combination of advanced microscopy, sequencing, and artificial intelligence allows us to elaborate on the initial plant cell atlas (Rhee et al., 2019). We also see great potential in complex studies and cell-based models describing morphogenetic processes.

This review aims to show how the combination of SC data, morphometric data, and cell-based models will expand our understanding of tissue and organ morphogenesis. We discuss the possibilities and prospects of such an integrative approach for solving reverse problems, including SC data and tissue imaging coupled with cell-based morphogenesis models. Finally, we consider available tools for cell-based models and present our cell-based modeling framework for morphogenetic processes. This algorithm is iterative and includes six main steps: (i) model formulation; (ii) design experiments to obtain microscopy and scRNA-seq data; (iii) obtaining experimental data; (iv) data analysis; (v) data integration into a hybrid (discrete-continuous) mathematical model of morphogenesis; (vi) model validation and verification.

2. EXISTING APPROACHES TO THE ANALYSIS OF SINGLE-CELL DATA AND THEIR POTENTIAL FOR CELL-BASED MODELS

Characterizing the plant cell fate and ontogenesis using SC technologies is a novel and promising approach for getting high-resolution genomic data that reveals new facts about various cell types. The first SC transcriptomic experiments have been carried out for the model plant *A. thaliana* in 2019. For *A. thaliana*, most of SC studies were conducted on root cells (Denyer et al., 2019; Jean-Baptiste et al., 2019; Ryu et al., 2019; Shulse et al., 2019; Turco et al., 2019; Zhang et al., 2019; Farmer et al., 2021). Whereas, there are only two studies conducted on leaf tissues (Kim et al., 2021; Lopez-Anido et al., 2021). Thus, for all the main cell types of roots and leaves, the developmental trajectories were revealed. Also, *Zea mays*, being a representative of C4-photosynthetic cereals, is a promising object for SC experiments due to the large size its cells, which allows to easily isolate specific cells, for example, from the shoot apical meristem. To date, there are studies based on the single-cell analysis for corn tissues carried out on a shoot apex (Satterlee et al., 2020), phloem (Bezrutczyk et al., 2021), and ears (Xu et al., 2021). The first and so far only scRNA-seq on rice roots (Liu et al., 2021) revealed significant differences in the characteristics of individual cell types in comparison to the cell types of *A. thaliana*, which indicates the presence of significant species-specific differences at the cellular level. A brief summary of the currently existing Sc-experiments is given in **Table 1**.

Abbreviations: *A. thaliana*, *Arabidopsis thaliana* L.; SC, single-cell; SCS, single-cell sequencing; scRNA-seq, single-cell RNA sequencing; RNA-seq, RNA sequencing; t-SNE, t-distributed Stochastic Neighbor Embedding; UMAP, Uniform Manifold Approximation and Projection; LSM, Laser Scanning Microscopy; LS, Light-Sheet Microscopy; SPM, Scanning Probe Microscopy; SIM, Structured Illumination Microscopy; 3D-SEM, 3-Dimensional Scanning Electron Microscopy; ODE, Ordinary Differential Equation; PDE, Partial Differential Equations.

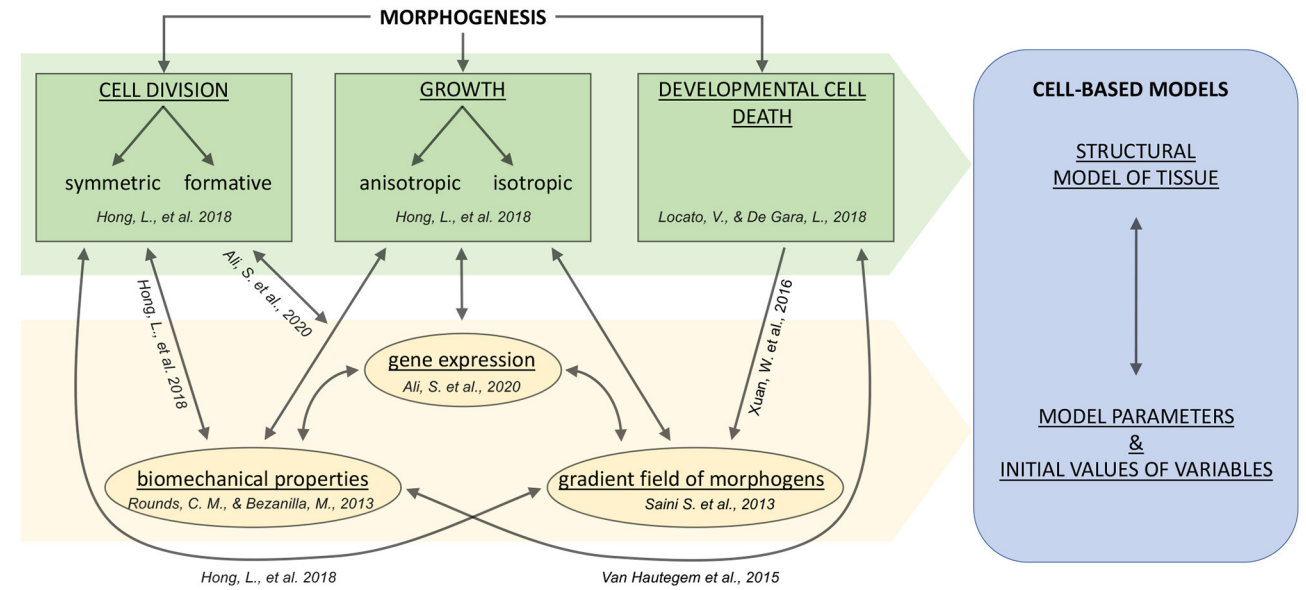


FIGURE 1 | A general scheme for systems biological and modeling concepts of plant tissue morphogenesis including cell growth and division, and developmental PCD (plant cell death). Arrows indicate the relationships between fundamental cell fate and intracellular processes. The cell fate processes are indicated in green; the intracellular processes or properties are indicated in yellow. The blue box indicates the significant components of the cell-based modeling approach. References correspond to theoretical articles briefly explained in the text.

TABLE 1 | Summary of scRNA-seq datasets obtained for plants.

Publication date	References	Drop-Seq platform	Illumina platform	Organism	Plant organ	Average reads per cell	Total genes detected	Expressed genes per cell
March 2019	Denyer et al., 2019	NanoDrop	NextSeq	*A. thaliana*	Root	87.000	17.000	4.276
April 2019	Ryu et al., 2019	10X Genomics	HiSeq 4000	*A. thaliana*	Root	75.000	22.000	5.000
May 2019	Zhang et al., 2019	10X Genomics	NovaSeq	*A. thaliana*	Root	40.000	23.161	1.875
May 2019	Shulse et al., 2019	Drop-seq v. 3.1	HiSeq 2500, HiSeq 4000, NextSeq	*A. thaliana*	Root	>1,000 UMI	20.464	1.549
May 2019	Jean-Baptiste et al., 2019	10X Genomics	NextSeq 500	*A. thaliana*	Root	19.000	22.000	2.445
July 2019	Turco et al., 2019	Drop-seq v. 3.1	NextSeq	*A. thaliana*	Root	NA	21.603	NA
April 2021	Lopez-Anido et al., 2021	10X Genomics	NextSeq500, HiSeq4000	*A. thaliana*	Leaf	70.000	NA	1.870
December 2020	Satterlee et al., 2020	Droplet microfluidics	NextSeq 500	*Zea mays*	Shoot	NA	NA	2000
January 2021	Kim et al., 2021	10X Genomics	HiSeq 2500	*A. thaliana*	Leaf	96.000	27.000	3.300
January 2021	Farmer et al., 2021	10X Genomics	HiSeq	*A. thaliana*	Root	NA	25.000	4.700
January 2021	Bezrutczyk et al., 2021	10X Genomics	HiSeq	*Zea mays*	Phloem	5,000	NA	NA
February 2021	Xu et al., 2021	10x Genomics	NextSeq 500	*Zea mays*	Ears	32.000	28.900	1800
March 2021	Liu et al., 2021	10x Genomics	HiSeq 2000	*Oryza sativa*	Roots	NA	NA	2600

There are several fundamental questions about the limitations and capabilities of the SC method (Rich-Griffin et al., 2020): How realistic is it to recreate a cell atlas using such data? Can we apply the technology to cells of any type? How to identify the main gene regulators and gene networks of development?

The problem of combining SC data from different plant species is of particular interest since the successful application of this approach can be used to create a unified developmental atlas. However, it is necessary to consider the species-specific features of tissue development and organization, which imposes certain restrictions on the joint interpretation of the exact SC data.

There is an acute lack of SC data of leaf and shoot stem cells except for *A. thaliana*. The small amount of existing SC transcriptome data is partly due to the complexity and length of the required experimentation and data analysis. In a recent overview of SC methods for plants (Lähnemann et al., 2020; Shaw et al., 2020), the authors highlight the major challenges and drawbacks of single-cell approaches: (i) gene expressing bias caused by the protoplasting procedure, (ii) unequal efficiency for extraction of different types of cells, (iii) difficulties for the reverse reconstruction of the cell atlas based on transcriptomic data, (iv) lack of data. We also want to point out that there are fuzzy boundaries between cell populations due to their connectivity and the presence of transport processes between them. Therefore, there are still several limitations to the biological interpretation of the SC data.

Thus, the classification of cell types and reverse spatial reconstruction are critical stages of SC transcriptome data analysis. This task is rather complex and requires using the original dimension of the expression data. SC data generally represent a filtered and normalized array with dimension $M \times N$, where M is the number of cells with a sufficient number of reads, N is the number of genes with a non-zero expression. The first component that can facilitate this problem is certain developmental trajectories caused by intracellular factors that limit the space of developmental possibilities and cause their partial determinism. Such factors have a different nature: the concentration of substances and energy substrates in the cell, the concentration of hormones and morphogens, the mechanical characteristics of cells (e.g., turgor pressure, tension, and thickness of the cell wall). Unfortunately, it is currently impossible to estimate the effect of these factors and their contribution to genes' expression. However, their presence makes it possible to identify the main differentiation genes. In general, this fact allows to carry out the procedure for reducing the dimensions of data. Depending on the data set's complexity, it is proposed to select from 1,000 to 5,000 highly variable genes for clustering and cell classification (Luecken and Theis, 2019).

A variety of available methods and tips for single-cell data dimensionality reduction and clustering are presented in the work of Nguyen and Holmes (2019). In most cases, researchers choose t-SNE and UMAP algorithms. The large computational complexity of the t-SNE method on big datasets was eliminated by adding fast Fourier transforms (Flt-SNE, Linderman et al., 2019). Comparison of t-SNE and UMAP methods revealed that UMAP outperforms even an optimized t-SNE in the computation time; also, clustering by UMAP is the most meaningful for distinguishing between cell types (Becht et al., 2019). Before the widespread use of t-SNE and UMAP, there was a probabilistic modeling method using Bayesian mixture of factor analyzers (MFA) (Campbell and Yau, 2017), based on the assumption that changes in gene expression are a linear function of time, which allows performing the Gibbs sampling procedure. This method's stability is inversely proportional to the number of genes with non-linear transient behavior, and its threshold was estimated in 40% of the total sample; if this threshold is exceeded, the authors recommend using the Diffusion Pseudotime (DPT) method (Haghverdi et al., 2016).

Also, machine learning demonstrates its consistency and efficiency in the analysis of SC transcriptomic data. For example, single-cell interpretation via multi-kernel learning algorithm (SIMLR) can perform dimension reduction, clustering, and visualization; this algorithm is characterized by enhanced performance and better visualization and interpretability compared to t-SNE, PCA, and zero-inflated factor analysis (ZIFA) methods (Wang et al., 2017). There are additional packages and algorithms for analyzing single-cell data, from preprocessing to data visualization; for example, on the Bioconductor platform (Amezquita et al., 2020), or the Python-based scalable toolkit SCANPY (Wolf et al., 2018).

Modeling the dynamics of gene networks is a promising approach for extracting biological facts from single-cell transcriptomics. When reconstructing such networks, it is possible to identify both transcriptional regulators and their targets. For example, a high-performance TENET protocol is based on the calculation of transfer entropy and can predict large-scale gene regulatory cascades and relationships in single-cell data (Kim et al., 2020). Also, there is SCENIC, a fast calculation Python algorithm that reconstructs the regulons (Van de Sande et al., 2020). Comparing the accuracy of calculations of gene networks by different algorithms showed that successful methods on artificial data sets are characterized by low accuracy on real data (Pratapa et al., 2020). The authors have selected three promising methods with high computational accuracy on real data: partial information decomposition and context (PIDC) (Chan et al., 2017), gene network inference with the ensemble of trees (GENIE3) (Irrthum et al., 2010), and GRNBoost2 (Moerman et al., 2019).

Elaboration of specific algorithms for using SC transcriptomic data to reconstruct developmental gene networks and identify new regulators remains a challenging issue. Databases and genetic interactions can serve as an additional source for expanding genetic networks and their verification. For example, STRING database (Szklarczyk et al., 2019) includes information about protein-protein interactions and allows to perform network reconstruction, visualization and functional enrichment analysis. Cytoscape is a suitable environment for further network visualization and addition of meta-information (Shannon et al., 2003). The functionality of this application has been significantly expanded due to the many available plugins. For example, the GeneMANIA plugin (Warde-Farley et al., 2010) allows to predict additional network elements and new connections, whereas the plugin *yFiles* (Wiese et al., 2004) provides additional tools for network layout.

Another ambitious challenge is the integration of multi-omics SC data. Ma et al. (2020) examines the capabilities

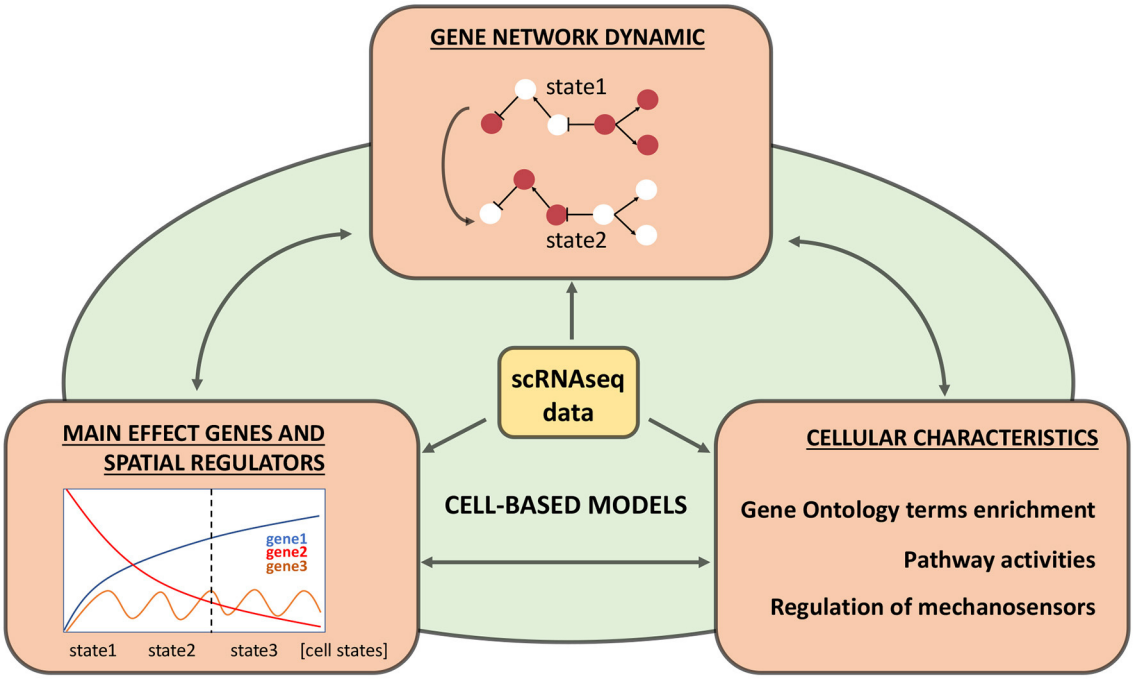


FIGURE 2 | Relevant information from single-cell transcriptomics experiments for cell-based models. Three types of information are highlighted in orange blocks, their integration into the cell-based model is shown in green, and double-headed arrows indicate each block's comparison. The central yellow block indicates original processed single-cell RNA sequencing (scRNA-seq) data.

of 10 SC integration tools and tests the functionality of the four most relevant ones (Giotto, MOFA, LIGER, Seurat3). It should be noted that the existing problems in the analysis and interpretation of data give rise to the rapid development of various methods and approaches to their processing. The available collection of various methods and tools for analyzing SC data is presented in this online repository. Also, pipelines and statistical methods useful for analyzing SC data are presented in the work by Petegrosso et al. (2020).

Although obtaining high-quality SC transcriptomic data for plants is a routine, standardized procedure, cell extraction processes, meaningful interpretation and verification of data are essential and non-trivial stages for the development of this technology. An important step in data validation and interpretation is the construction of mathematical cell-based models, which combines the data about concentration of morphogens and expression of genetic regulators inside the cells and "rules," which determine intercellular communications, cellular mechanics, transport processes as well as the transition between cellular states. However, with current technology, we cannot directly use the entire array of transcriptome data to create mathematical models of morphogenesis due to the large number of dimensions. Therefore, it is important when comparing different cell types to identify the main genetic and metabolic differences and take them into account in models.

There are a few methods, which can potentially allow researchers to use scRNA-seq data for building the cell-based models (see **Figure 2**):

1. Identifying crucial genes (main effect genes) and regulators which explain a lot of variance/differences between cell types.
2. Searching for novel regulatory genes, which have a spatial distribution of expression between cells of different types.
3. Reconstructing Boolean gene networks using transcriptomic data.
4. Estimation of differences in integral characteristics (such as biomass, wall thickness, concentration of metabolites).

For example, SC transcriptome data could provide some indirect estimations of the cell wall's mechanical properties. The main mechanosensing genes are described in Du and Jiao (2020): receptor-like kinase FERONIA (FER), Leucine-rich repeat extensins (LRXs), DEFECTIVE KERNEL 1 (DEK1), and their targets of cell wall integrity pathways. Therefore, assessing these genes' expression levels in different cell types can potentially describe their mechanosensitivity and cell wall stiffness. Thus, SC data allows the definition of cell types' molecular characteristics, identifies regulatory subnetworks, and assesses their dynamics. These data can potentially be taken into account as parameters in cell-oriented models.

3. MODERN IMAGING TECHNOLOGIES FOR OBTAINING DATA ON PLANT TISSUES WITH A SINGLE-CELL RESOLUTION

Spatial organization plays a significant role in each cell's fate, affects transport, the direction of division, apoptosis, and the

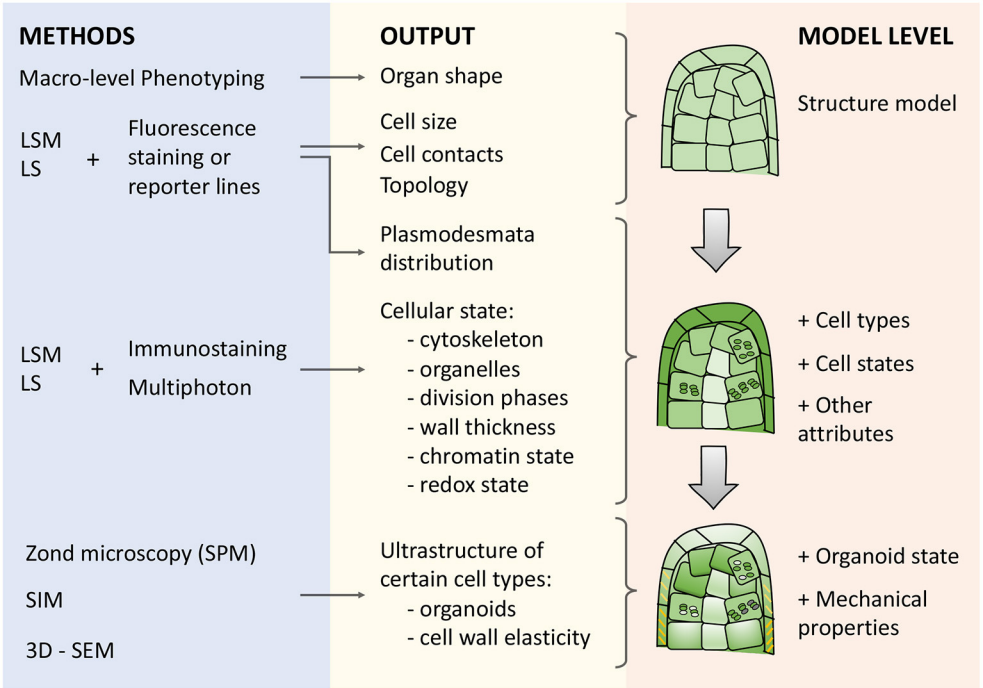


FIGURE 3 | Types of microscopy techniques, their outputs, and meanings for describing morphogenetic processes in cell-based models. There are three blocks in the scheme: (i) methods (blue box), (ii) corresponding outputs (yellow box), and (iii) model levels (orange box) from structural to organoid resolution. Abbreviations used: LSM (Laser Scanning Microscopy), LS (Light-Sheet microscopy), SPM (Scanning probe microscopy), SIM (Structured Illumination Microscopy), 3D-SEM(3-Dimensional Scanning Electron Microscopy).

cells' structural peculiarities. Therefore, this information is the basis for a systemic integrative study of the processes of morphogenesis.

The cells of vascular plants form a shared symplast through the cell walls, which determine the fixed position of the cells in the tissue (Vaahtera et al., 2019). In plants, cell migration is almost absent, but in some cases, cells can shift their positions relative to each other: part of the plant cell remains in its original place, while other parts of the cell grow to the new locations, moving significantly relative to other cells (Lev-Yadun, 2015).

There are various specialized approaches for phenotyping (**Figure 3**): visible light, spectroscopy, infrared, fluorescence, 3D, and tomographic methods for getting plant images (Li et al., 2014). The imaging techniques for plant quantification are broadly used due to their inexpensive cost, simplicity of operation, and maintenance (Omari et al., 2020).

Reconstruction of plant architecture in terms of shape, size, and topology of cell connections (**Figure 3**) is an essential component to reach an integrative systemic understanding of aspects of the functioning of both individual cells and tissue as a whole (Fricker, 2016; Zubairova et al., 2019; Kerstens et al., 2020). A variety of optical tissue imaging techniques (**Figure 3**) currently allow access to such cellular characteristics (optical and fluorescent microscopy, laser scanning approaches, and structured lighting microscopy). Since higher plants' organs are multilayered and volumetric, imaging techniques based on 3D analysis of a fluorescent signal, such as laser scanning microscopy, are currently among the most widespread visualization methods of cellular architecture. It allows to reconstruct the architecture of tissue and organ fragments consisting of thousands of cells (Zubairova et al., 2019) and to analyze *in vivo* large time-series for reconstructing the dynamics of development (Goh, 2019; Seerangan et al., 2020).

Together with modern image analysis methods, they provide a reliable decomposition of cell layers and assessment of cell morphological parameters (Legland et al., 2016; Erguvan et al., 2019; Zubairova et al., 2019). The number of cells reconstructed by ImageJ-plugins LSM-W^2 (Zubairova et al., 2019), SurfCut (Erguvan et al., 2019), as well as MorphoGraphX instruments (Kerstens et al., 2020) is limited by the computer performance and technical capabilities of the microscope. They allow working on a local computer with arrays from thousands of cells, which is of a comparable order to scRNA-seq methods. The most comprehensive range of methods makes it possible to segment cells, measure cell shape parameters, and reveal the topology of cells' connection with each other (Jackson et al., 2017).

Over the past few years, the possibility to study many entire organs through complete reconstruction at the cellular level became a significant breakthrough (Wolny et al., 2020). The root tip of *A. thaliana* is the most abundant target for scRNA-seq in plants. At the same time there are many reconstructions and 3D atlases for it (Dolan et al., 1993; Bowman, 2012; Mai et al., 2014) and even specialized software that allows displaying the various cellular characteristics into cellular ensembles, for

example, the iRoCS Toolbox (Schmidt et al., 2014). *In vivo* laser scanning microscopy techniques coupled with mathematical modeling allowed describing the processes of morphogenesis for the arabidopsis root apical meristems (Mironova et al., 2012). The dynamics of the development of *A. thaliana* lateral roots are also available for visualization at the cellular level from the earliest stages of their establishment (Goh, 2019). Using confocal and multiphoton microscopy approaches, apexes and leaf primordia can also be completely reconstructed (Kiss et al., 2017; Wolny et al., 2020), as well as adult leaves (Wuyts et al., 2010) and sepals (Tauriello et al., 2015).

3D reconstruction of *A. thaliana* ovule coupled with transcriptome sequencing provides incredibly detailed data about developmental processes of this organ (Vijayan et al., 2021), which can serve as a set of reference points for further integration of future single-cell data on this organ. Simultaneously, the methods of visualization and analysis of images also allow working with plants with larger organs, for example, with *Nicotiana tabacum* roots (Pasternak et al., 2017).

Light-sheet imaging techniques allow to increase the scan depth and improve the quality of the reconstruction. These technologies, coupled with mathematical modeling, gave insights into the geometrical organization of divisions during the formation of the lateral root of *A. thaliana* (von Wangenheim et al., 2016). In particular, the first division of the cell-founders is always asymmetric and determines the formation of a layered structure, while the pattern of further cell division forms thanks to a regular change in the orientation of the division plane. Also, the technique of optical cleaning of plant tissues allows for getting deep 3D imaging and is compatible with fluorescence-based microscopy (Warner et al., 2014). The measurements of morphological characteristics of cells and their mutual arrangement allowed researchers to form a structural model of the studied organ and identify cell types (Kerstens et al., 2020).

The current opinion about coordination of growth processes and divisions (Sablowski, 2016) stressed the role of individual cell characteristics and intercellular interactions in these processes. Optical microscopy is a valuable method for obtaining the structural characteristics on the subcellular resolution. For example, this approach allows studying the ultrastructural features of the cell wall (Yarbrough et al., 2009), which enables us to assess cellular biomechanics indirectly. The combination of large-scale annotated image datasets and deep learning approaches is a promising technique for annotating physical, morphological, and tissue grading cellular properties (Fricker, 2016; Biswas and Barma, 2020).

The cell wall's mechanical parameters deserve special attention since they determine features of the growth process (Bidhendi and Geitmann, 2016), and therefore is incredibly important for modeling plant morphogenetic systems. In addition to assessing the thickness of the cell wall (Krzesłowska et al., 2019), modern approaches make it possible to evaluate its composition and mechanical parameters. For example, probe microscopy can assess the spatial composition of polysaccharide filaments on the surface of living tissues (Zhang et al., 2016), and Raman microscopy can produce data on the composition and ultrastructure of the cell wall on sections of organs in the usual (Zeise et al., 2018) and confocal modes (Gierlinger et al., 2012). The ultrastructure of cell walls as well as tissues and organs can be studied with a 3D electron microscope (Kremer et al., 2015). All these methods make it possible to assess biomechanical parameters within organs and serve as the basis to improve the simulation modeling of growth processes.

Therefore, the next important step is integrating the structure model with the cell parameters that mark the individual and group characteristics of cells (**Figure 3**). Many characteristics of the nucleus, organelles, and cell walls can be identified at the scale of an entire organ using approaches of protein immunolocalization, expression of reporter constructs that mark certain cellular features, as well as using methods to increase the resolution of microscopy (**Figure 3**).

The data on the frequency of mitoses along the root (Pasternak et al., 2017; Lavrekha et al., 2020) provides insight into the dynamics of replenishment of cell files and the size zones, where cell divisions occur. Also, cells in S-phase can be identified by incorporating labeled nucleotide analogs (Pasternak et al., 2017). The passage of the cell cycle phases is closely associated with the cell fate specification (Roeder et al., 2012). The state of chromatin in cells of various types can be identified using immunolocalization (She et al., 2018) and shed light on cell activity. Visualization of the cytoskeleton can be done both by immunolocalization, staining with phalloidin, and, *in vivo*, using reporter genetic constructs (Zhang et al., 2020). These cells' characteristics can be related to changes in gene groups' expression in cells and are suitable for improving the integration of the structural model with single-cell transcriptomic data.

The distribution of various proteins in plant organ cells can also be determined (Sauer and Friml, 2010) and used for integration into a model. Proteins can be transporters that determine the fluxes of substances that deserve special attention; for example, the auxin membrane transporter PIN1 has a significantly uneven distribution over root cells and a polar arrangement on the cell surface (Omelyanchuk et al., 2016). It has also been shown that RNA molecules capable of being transported from tissue to tissue play an essential role in the regulation of biological processes in a plant, and their visualization within an organ is also possible (Luo et al., 2018).

Also, plasmodesmata play a unique role in the processes of intercellular symplastic transport and signaling in plant tissues (for comprehensive review, see Heinlein and Epel, 2004). Plasmodesmata are intercellular channels characterized by various states from open to closed (Crawford and Zambryski, 2001). Plasmodesmata behavior underlies the isolation of groups of cells in the tissue, called symplastic domains (Pfluger and Zambryski, 2001; Lucas and Lee, 2004; Yadav et al., 2014). Stress factors affect the formation of plasmodesmata (Fitzgibbon et al., 2013). The transport of mRNA and metabolites through the plasmodesmata affects the concentration of substances and gene expression levels inside particular cells (Lucas and Lee, 2004). Many non-cell-autonomous transcription factors and small RNAs are known to move through plasmodesmata between cells and regulate their interaction during development (Kragler, 2013; Yadav et al., 2014; Sevilem et al., 2015).

Transmission electron microscopy is the classical method for studying the morphology of plasmodesmata. Combined with light-based microscopy, it allows one to study the structure and distribution of plasmodesmata between cells of specific cell types (Nicolas et al., 2017). Also, the topology of plasmodesmata of contacting cells at organ scale can be studied using confocal and super-resolution microscopy (Fitzgibbon et al., 2010, 2013). In this sense, microscopy allows us to assess the location and topology of plasmodesmata and, therefore, identify the potential of local transport of substances through these transport channels, symplastic domains, and to assess the order of cell division. Thus, the organization and localization of transport channels inside the plant tissues are connected with the intracellular characteristics.

On the other hand, intracellular sensing processes contribute to intercellular signaling. For instance, there are special sensory plastids in epidermal and vascular parenchyma cells, which can cause a global systemic stress response in a plant (Beltrán et al., 2018).

The redox state of organelles is also an additional factor associated with developmental processes, ROS signaling, and antioxidant systemic plant cells (Bobrovskikh et al., 2020). In particular, the CellROX fluorescent reagent visualizes the oxidative potential of cells in a tissue (Kováčik and Babula, 2017).

Besides, mass spectrometry imaging and live single–cell mass spectrometry practically corresponds to single-cell metabolomics and makes it possible, for example, to mark the concentrations of secondary metabolites on the whole adult organ (Yamamoto et al., 2019). Such approaches can be combined with SC analysis of the expression of these metabolites' biosynthetic enzymes and transporters. As a result, they provide a basis for modeling the distributed regulation of these processes at the tissue level (**Figure 3**). The most important polynucleotides, such as RNA, can also be detected at the level of single molecules (Huang et al., 2020), which allows direct integration into the structural model of the organ.

Modern imaging techniques allow access to the structural and physiological characteristics of cells in a whole organ manner. It provides ample opportunities to create, enrich, and verify structural models of plant organs and tissues. An important aspect is that many assessments can be carried out over time. Comparison of temporal dynamics in zones with active morphogenetic events will make it possible to track changes in cellular topology, and thus, to trace the nature of division (symmetric and asymmetric) and growth (isotropic and anisotropic), as well as to detect several mechanical features of the developing tissue (for example, the relative stiffness of different cell zones).

Thus, a large arsenal of available microscopic and imaging techniques allows obtaining high-quality multilevel data integrated into plant morphogenesis models. For example, there is a computational morphodynamics approach that allows formalizing quantitative data from morphometry measurements into a set of rules (Formosa-Jordan et al., 2018):

1. To set ODE, which describes the growth rate of individual cells using data from regulatory networks.
2. To set various rules for the geometry of division (periclinal/tangential divisions with different angles) according to mechanical constraints of intercellular vertex interactions.
3. To use the first two steps to calculate effective growth and final rate equation.

4. CELL-BASED MODELING APPROACHES REPRODUCING PLANT TISSUE MORPHOGENETIC PROCESSES

4.1. Existing Models and Modeling Approaches

This section will discuss existing mathematical models describing the tissue organization and/or properties of individual cell types. While considering plant growth and developmental processes, researchers often highlight a unique role for the hormone auxin. For instance, in plant roots, auxin triggers cascades of events during development and morphogenesis, while other hormones (cytokinins, brassinosteroids, abscisic acid, gibberellins, and others) interact with auxin (Saini et al., 2013). Auxin is also an important regulator in developing shoot apical meristems in combination with cytokinins, gibberellic acid, and some transcriptional factors: WUSCHEL, ARR7/ARR15, ARF5 (Durbak et al., 2012). Mironova et al. (2012) demonstrated the effectiveness of the reverse fountain and the reflected flow mechanisms of PIN-associated transport in the root apical meristem. Comparison of different complexity models showed that a model that only describes auxin transport processes is insufficient for the reproduction of realistic patterns of morphogenesis, but adding an additional layer-specific regulation or layer-driven growth could help solve this problem (De Vos et al., 2014).

Simultaneously, the mechanical characteristics of tissues, which are determined through a complex interplay of genetic and physiological systems, are an essential component for describing the processes of morphogenesis. The feedback effects of mechanical interactions and stresses, which affect the regulation of proliferation patterns, are highlighted in Nelson et al. (2005). The experimental evidence of the mechanical stress approach's consistency for plant tissue development is shown in the work of Uyttewaal et al. (2012). The transition from the linear models of hormonal transport to hybrid multicellular and multiscale models has excellent potential for predicting the emergent properties of the system (Voß et al., 2014). The basis for mechanical models of cell growth is the representation of multicellular tissues in vertex-based graphs with the calculation of the interaction forces between these elements. The equations binding the growth of plant cells with the rate of water absorption and the cell wall's growth were first published in Lockhart's work for the case of constant turgor pressure (Lockhart, 1965). In order to model growth in a more general case, Lockhart's equations were extended, taking into account the change in turgor pressure as a result of reversible elastic deformation and transpiration processes in the Ortega model (Ortega, 2010). Within the framework of this approach, a linear leaf growth model was proposed (Zubairova et al., 2016). In addition, Newton's First Law and Hooke's Law can be used to describe

TABLE 2 | The most popular tools for cell-based plant tissue morphogenesis modeling.

Name, reference, link	Spatial scale	Formalism	Examples
Virtual cell (Moraru et al., 2008)	2D/3D	Kinetics, diffusion, flow, membrane transport, electrophysiology	Gajdanowicz et al., 2011; Onal et al., 2020
OpenAlea (Pradal et al., 2008)	2D/3D	Functional-structural plant models	Muraro et al., 2014
CellModeller (Dupuy et al., 2008)	2D	Biphasic systems; viscous yielding of the cell walls	Dupuy et al., 2010; Rudge et al., 2012
VirtualLeaf (Merks et al., 2011)	2D	Vertex dynamics model	van Mourik et al., 2012; De Rybel et al., 2014; De Vos et al., 2014
CompuCell3D (Swat et al., 2012)	2D/3D	Cellular Potts model	Hester et al., 2011; Swat et al., 2015
CellZilla (Shapiro et al., 2013)	2D	Vertex dynamics model	Nikolaev et al., 2013; Shapiro et al., 2015
LBIBCell (Tanaka et al., 2015)	3D	Lattice Boltzmann method for solving fluid and signaling processes	Stopka et al., 2019

cell growth and expansion, as was done in the recent work by Retta et al. (2020).

Unfortunately, most available auxin-related models are focused only on the transport processes in the root tissue and poorly explain the overall processes of growth and development (Morales-Tapia and Cruz-Ramírez, 2016). However, several models combine both a mechanical approach and auxin transport processes. For example, there is a dynamic model that describes molecular mechanisms in conjunction with physical tension fields and auxin dynamics (Barrio et al., 2013). This model reproduces emergent patterns of morphogenesis from proliferative to transition and elongation zones. The study combining experimental data on the organization of the extracellular matrix and numerical simulations demonstrated that auxin plays an essential role in altering cells' mechanical properties; this process involves the ABP1 and KATANIN 1 proteins (Sassi et al., 2014). Also, the advanced cell-based mathematical model describes the relationship between the concentration of morphogens and the cellular mechanistic properties in the developing apical shoot meristems (Banwarth-Kuhn et al., 2019).

Thus, the models of plant tissue morphogenesis put at the forefront three biological facts: (i) the dependence on intercellular hormonal signaling, (ii) the importance of the intracellular state and individual cellular characteristics, (iii) the relevance of mechanical stresses in intercellular interactions. Therefore, scRNA-seq technologies, microscopy, imaging techniques, and a range of complementary approaches to measuring cell mechanical properties (Banwarth-Kuhn et al., 2019; Bidhendi and Geitmann, 2019) can provide a complete picture of morphogenetic processes at the cellular level.

4.2. Available Software and Tools for Cell-Based Modeling

In general, elaborating mathematical models of morphogenetic processes could base on specialized software, which we discuss in this section. Researchers may also develop and implement their frameworks and algorithms using mathematical packages and general-purpose programming languages (Python, Mathematica, MATLAB). Three formalisms are most often used to build cell-based models: vertex-based, center-based (also called spring-based), and Cellular Potts models. Vertex-based models are often used to simulate plant tissue and make it possible to conveniently describe the dynamics of cell movements in cell ensembles taking into account mechanical constraints (for example, during morphogenesis). This formalism is implemented in the Cellzilla (Shapiro et al., 2013), VirtualLeaf (Merks et al., 2011) packages. In center-based models, cells are represented as dots with mass, connected by mechanical elements (springs). Banwarth-Kuhn et al. (2019) give an example of this formalism's application to the description of growth processes in the shoot apical meristem. Cellular Potts models are often used to describe the processes occurring in animal tissues and tumor formation processes; this formalism is implemented in CompuCell3D (Swat et al., 2012). It is also possible to use the Voronoi tessellation formalism for modeling morphogenetic processes; e.g., see Romero-Arias et al. (2017).

Below we discuss available software, while a summary is presented in **Table 2**; for more details, see **Supplementary Table 1**.

Virtual Cell (Cowan et al., 2012; vcell.org) is an environment for modeling, analysis, and simulation of cellular processes, and it includes tools for gene network and for the integration of biological images. This package consists of distinct functional modules: rule-based networks, ODE, PDE and kinematics, stochastic simulations, parameter estimation and has the ability to integrate it into hybrid models. Users can define the model structure and the system automatically builds the code and compiles it. A detailed overview of this tool is given in Moraru et al. (2008). Also, there is a VCell extension for compartmental and spatial rule-based modeling (Blinov et al., 2017). The implemented models using VCell can have a different scale, for example, the model of potassium transport in plant vascular tissues (Gajdanowicz et al., 2011), and model of the paracrine-juxtacrine loop for breast cancer cells and macrophages (Onal et al., 2020).

VirtualLeaf package (code.google.com/archive/p/virtualleaf/, Merks et al., 2011) using a vertex-based approach (Nagai and

Honda, 2001); the algorithm includes vertex motions at each step that minimize the Hamiltonian energy by the Monte Carlo algorithm. For each cell, an unstressed area is specified, corresponding to the cell's state when the turgor pressure is balanced with the external pressure. For each cell wall element, the unstressed length is specified, corresponding to the length of the cell wall segment in the absence of turgor pressure. The balance between turgor pressure and the cell wall's resistance can be described in terms of the generalized potential energy (Hamiltonian) calculated as the sum of all cells and cell wall elements, which is then minimized by the algorithm. The growth models of root were implemented using this framework (De Vos et al., 2014).

Cellzilla uses a vertex dynamics model for describing morphodynamics processes and takes into account morphogenetic regulation (http://cellzilla.info/, Shapiro et al., 2013). The cellular structure is represented by a list of three elements: a list of vertex coordinates, a list of edges consisting of pairs of vertex numbers, and a list of cells consisting of lists of edge numbers belonging to a cell. The interaction between morphogens and the transport flows in each cell is described in terms of chemical kinetics using the arrow notation of the Cellerator package (Shapiro et al., 2003). This software automatically constructs and solves a system of differential equations describing the dynamics of morphogens' concentration in all tissue cells. Methods for constructing models of plant cell growth in CellZilla are described by Shapiro et al. (2013). Using this system, Nikolaev et al. (2013) constructed a model for *A. thaliana* shoot apical meristem structure maintenance.

CellModeller (haselofflab.github.io/CellModeller/; Dupuy et al., 2010) is a software with modular structure for 2-dimensional simulations. It can reproduce the intracellular dynamics of metabolites, intercellular transport processes, as well as cell mechanics using physical laws. This software can be used for modeling plant morphogenetic processes. For example, a simple morphogenetic system for the Coleochaete alga has been developed (Dupuy et al., 2010).

LBIBCell (Tanaka et al., 2015, https://tanakas.bitbucket.io/lbibcell/) was developed specifically to simulate morphogenetic processes in tissues. This tool uses the immersed-boundary concept (which describes cells as viscous fluid with elastic walls), coupled with the Lattice Boltzmann method. The model of biased epithelial lung growth was implemented using this tool (Stopka et al., 2019).

OpenAlea (Pradal et al., 2008) is an integrative platform that combines various computational frameworks. This platform's main goal is the integration and mutual enrichment of experience in different sections of plant process modeling. This system is based on Python language and has a visual programming interface. For example, the OpenAlea package VPlants (https://team.inria.fr/virtualplants/) allows building models of tissue morphogenesis. This package was used in modeling vascular development in *A. thaliana* (Muraro et al., 2014).

CompuCell3D (Swat et al., 2012) is a C++ software for 3D modeling, which includes both graphical user and command-line interfaces. This system uses classical mechanics for describing cellular behavior according to mechanical constraints. Multicellular systems are described using the Cellular Potts model. The input data include the grid's size, number of cells, cellular interactions, energy functions, and activator concentrations. The protocol for using this program to study cellular morphogenesis parameters is presented in Palm and Merks (2015). Most of the models elaborated with this software describe the development of animal tissues (Hester et al., 2011) and the processes of tumorigenesis (Swat et al., 2015).

Thus, the available software and methods are pretty diverse, and the choice of a particular tool depends on the specifics of the task at hand. Among these tools, it is necessary to highlight Cellzilla and VirtualLeaf as the most specific for describing plant morphogenesis processes. On the other hand, the development of new frameworks and algorithms, which depend on researchers' ability to program, is a promising approach since it significantly expands the functionality and removes several restrictions on applying one or another formalism implemented in existing software.

4.3. Our Framework and Model Flowchart

In this section, we propose a general framework for modeling plant morphogenetic processes based on various biological data. This kind of model should include two main data sources: scRNA-seq and tissue imaging data; besides, SC metabolomics and cell wall stiffness studies can serve as additional data sources. For plant organ growth modeling, the accurate description of processes on the cellular level is essential since this level combines molecular regulation with hormonal regulation, cell division, and reproduction processes (De Vos et al., 2012).

Mathematically, events occurring in plant tissues and cells can be classified into continuous and discrete ones. The first ones include the processes of metabolism, growth, transport and development of cells. Discrete events, on the other hand, include processes such as birth (or emergence), division, death, and change of cellular state. Individual cells' metabolic characteristics are influenced by their genotype and developmental stage, which would be described by single-cell transcriptomics approaches. The nature of the proposed framework is hybrid since it combines different mathematical formalisms and modules: (i) ODE/PDE equations for describing the dynamics of substances and morphogens inside the cell and the processes of intercellular transport, (ii) discrete events occurring during the onset of threshold conditions (for example, cell division when a specific cell area is reached, or cell differentiation at a hormone concentration above the threshold), (iii) the biophysical laws of mechanical interactions between cells (such as Ortega's approach Ortega, 2010 or Newton's and Hooke's laws Retta et al., 2020). In this sense, scRNA-seq data helps measure individual characteristics of cell populations (which characterize system dynamics), while microscopy should help to define geometrical patterns and "rules" (e.g., division geometry or dividing plane orientation). These steps will help to create hybrid models with tissue/cellular resolutions.

The usefulness of such a hybrid approach in describing ecological systems was described in the work of Vincenot et al. (2011). In particular, the combination of discrete and continuous

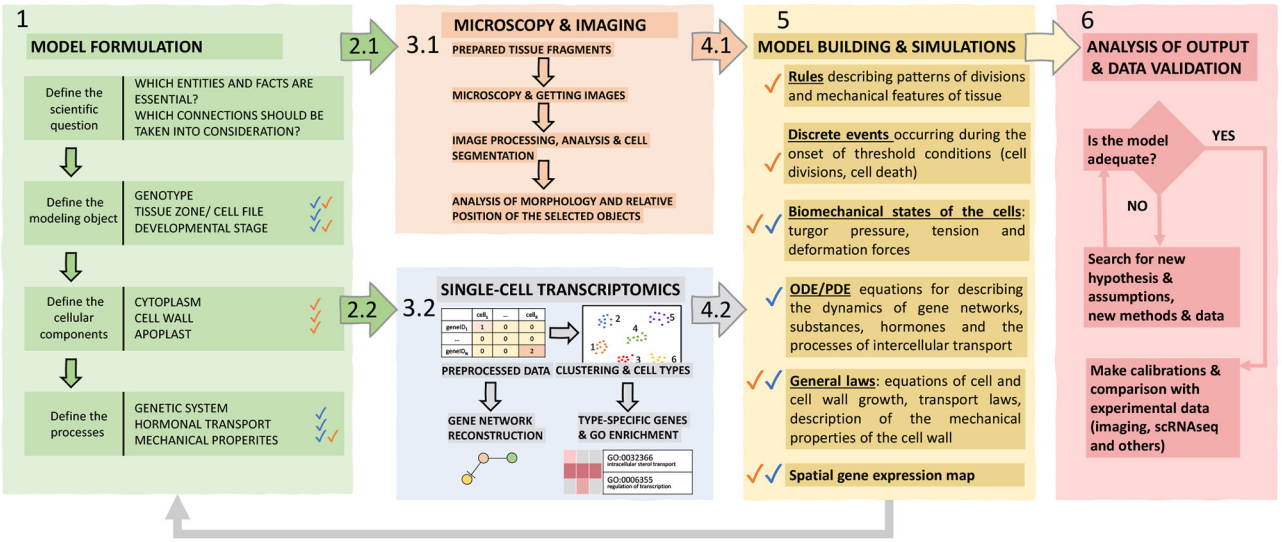


FIGURE 4 | The proposed hybrid framework for cell-based models construction. The framework includes six functional blocks explained in the text. Individual blocks are marked with corresponding colors; colored arrows indicate the transition between blocks. Blue and orange checkmarks indicate information related to single-cell and imaging data, respectively.

phenomena is a natural property of multicellular systems, and such hybrid frameworks allow researchers to make more realistic simulations *in silico*. Van Liedekerke et al. (2015) described the advantages and disadvantages for different types of agent-based models of tissue mechanics and noted that hybrid models could reproduce spatial resolution, physical aspects of interactions, cell shapes diversity. Osborne et al. (2017) compared different approaches to cell-based modeling using typical cases of the described processes; the authors noted that the vertex-based approach, in contrast to others, allows one to simulate boundary conditions in proliferation processes effectively. This feature allows us to consider this method as the most promising for modeling the root apical meristems, which has more severe mechanical restrictions for growth than leaf and shoot tissues. For modeling leaf and shoot tissues, for example, it is possible to use the Voronoi tessellation or overlapping spheres modeling approach described in Osborne et al. (2017).

Thereby, we assume the use of such a hybrid approach complementary to modern research due to its multilevel nature; it combines SC transcriptomic and microscopy data into a cell-based modeling framework. Below in the text and in **Figure 4**, we outline the main stages of our framework that must be taken into account.

1. The posed biological problem determines the structure of the model. A modeler should define a biological system's properties, its elementary subsystems, and connections between these elements, which are significant to reproduce them in the model. Based on these decisions, it is necessary to determine the main properties of the simulated object: genotype, organ, tissue zone, stage of development. Since a cell is a crucial element for describing the processes of plant morphogenesis, the next step is to find out which cellular structures will be reproduced in the model to determine the formalism used to describe them and the equations for growth and the rules of division. Then, it is necessary to decide on the objects at the molecular level to be considered, in particular the genetic systems of interest, to find out whether it is required to consider transport processes for morphogens (for example, hormones), and also to decide whether it is necessary to take into account the biomechanics of cells for the modeled system.
2. Designing experiments to obtain imaging (2.1) and scRNA-seq (2.2) data based on the given aim. For imaging (2.1), it is essential to choose a suitable plant portion and microscopy technology and determine whether it is necessary to track the dynamics of development of a given fragment of tissue and for which interval of time. For scRNA-seq (2.2), it is important to make sure that the process of isolation of protoplasts and their analysis will not be limited due to the structure of the tissue and/or organ of the plant, imperfections, and shortcomings of the available methods, otherwise, this technique will have to be worked out and improved to an acceptable level.
3. Perform the experiments and produce data. (3.1) It is necessary to prepare (for example, fix and stain) a target tissue fragment, get images, process and analyze them (manually or using plugins), and digitize the resulting patterns to build a structural model of the tissue/organ and identify morphogenetic rules for incorporation into a computational model. (3.2) While obtaining and analyzing scRNA-seq data, special care should be taken to ensure that the research aim is as close as possible to the intended modeling goals. Care should be taken to avoid contamination with cells of those classes that are not needed and so that for most of the required cells, it would be possible to analyze the molecular systems required for the model. Besides, scRNA-seq-based approaches

for the reconstruction of gene networks of the corresponding processes have high potential.
4. Analyze experimental data. Experimental results at cell and tissue level have to be analyzed in order to derive key parameters to be used in the model formulation in terms of cellular characteristics (4.1) and molecular processes (4.2) for all the considered cell types.
5. Systematic assembly of the hypotheses, available data and mathematical formalization into a single hybrid model, which consists of the following blocks: (1) ODE / PDE equations for describing the dynamics of substances and morphogens inside the cell and the processes of intercellular transport, (2) discrete events occurring at the onset of threshold conditions (for example, cell division when a specific cell area is reached, or cell differentiation at a hormone concentration above the threshold), (3) biomechanics interactions between cells (4) agent-based rules describing patterns of divisions and mechanical features of the tissue.
6. Validation and verification of models is based on their success in reproducing the behavior of real biological phenomena that can be evaluated experimentally. In this sense, it can be useful to return to the stage of morphometry and compare the dynamics of tissue development with simulations and study in detail the molecular organization of the subsystems described in the model.

In general, the proposed approach is universal for describing any morphogenetic system; however, the pipeline described above may differ in some steps for each specific case, while some of them could be eliminated. In particular, plant tissue morphodynamics is context-dependent due to mechanical interactions inside cell ensembles and the transport of morphogens through plasmodesmata, which is confirmed by numerous studies (Crawford and Zambryski, 2001; Heinlein and Epel, 2004; Lucas and Lee, 2004; Kragler, 2013; Yadav et al., 2014; Sevilem et al., 2015; Luo et al., 2018). At the same time, models for morphodynamics of animal tissues with strong neighborhood structures could include analogous mechanisms modified to consider cell adhesion processes. For example, this approach is applicable to model the processes of animal epithelial or tumor growth (Interian et al., 2017).

5. CONCLUSIONS AND FUTURE CHALLENGES

Post-genomic technologies made it possible to obtain detailed information about processes at genomic and transcriptomic levels using SC and whole tissue RNA sequencing technologies. Besides, the existing abundance of microscopy methods allows high-quality characterization of morphology and physiology at the level of extended fragments of tissues and organs. However, microscopy approaches do not allow to perform quantitative assessments of important intracellular characteristics, such as concentrations of substances and metabolites. SC metabolomics approaches for plants, which are beyond this review's scope, still remain overshadowed, although significant developments have been made in mass spectrometry approaches for such kind of analyses (de Souza et al., 2020). Gilmore et al. (2019) discuss the latest advances in mass spectrometry imaging: matrix laser desorption ionization (MALDI) and secondary ion mass spectrometry (SIMS), which have a high potential for assessment of metabolism at subcellular spatial resolution. The development of these methods will allow metabolomics to achieve the same spatial resolution level as SC transcriptomic. The review of Bidhendi and Geitmann (2019) presents the main features and possibilities of measuring the cell wall's mechanical properties: indentation technique, tensile test, acoustic microscopy, fracture measurements, and microfluidics. The authors emphasize that multiscale *in silico* mechanical modeling has excellent potential for the field and could help obtain a unified understanding of mechanical behavior across different scales.

To date, the methods and technologies necessary to obtain various experimental data for plant morphogenesis models have reached a balance and are mostly consistent with each other in terms of power, productivity, and spatial resolution. The community of mathematical biologists and programmers faces crucial theoretical challenges and is creating efficient computational frameworks capable of large-scale numerical simulations involving cellular ensembles of several thousands of cells. Such models will provide more accurate resolution and realism in the description of morphogenetic processes. Examples of optimization works are the algorithm of Jeannin-Girardon et al. (2015), and graphics processing units (GPU) accelerated framework for 3D cellular growth and division models (Madhikar et al., 2018). Moreover, declarative modeling perspectives concerning morphogenetic processes are considered (Mjolsness, 2019), which potentially will help formalize mathematical calculations at higher levels compared to general-purpose programming languages.

The widespread development of SC technologies in the future could serve as a driver for other areas of cellular and developmental biology of plants (Libault et al., 2017). However, we have an urgent need for data integration to successfully apply the technology, in particular at tissue level with its organization's peculiarities as an emerging system. Besides, an increased availability of SC data can stimulate the development of methods and modeling concepts at cellular and tissue levels, which will open the way for the binding of multi-omics characteristics for individual cell types and the observed phenotype.

On the other hand, it is necessary to verify the emerging issues related to the interpretation and analysis of SC data using advanced microscopy and *in silico* biology. In this sense, one of the most urgent problems of SC sequencing is the reverse reconstruction of the spatial position of cells based on corresponding transcriptome expression. Searching for major regulatory genes that characterize certain cell lines will be a critical step to solve this problem. Also, cell-based models of morphogenesis could help interpret and integrate SC and imaging data, making the reasoning more transparent and establishing an understanding of essential parameters and mechanisms for the described systems.

Summarizing all of the above, we have found the following key features related to SC-technologies that need to be addressed:

1. Some limitations are still present in the phases of integration, analysis, and interpretation of data.

2. Only a limited set of plant species and organs is suitable for obtaining transcriptome and structural data with cellular resolution.
3. There is a need for a more precise reconstruction of scRNA plant atlases.

The task of elaborating and analyzing *in silico* models of morphogenesis, due to the complexity of the studied systems and computational limitations, are non-trivial. Thus, cell-based models, which use a hybrid formalism, could effectively combine our knowledge on different levels and help tackle the complexity of the system. However, the current problem of the large number of dimensions of the initial SC data should be solved by applying preprocessing and filtering algorithms, as well as for the reconstruction of related gene networks. Thereby, model formulation and numerical experiments *in silico* could be applied using only the essential part of the initial high-dimensional SC data. Such reduction should aim to contain data on gene expression changes and metabolites concentrations, which determine the different cellular states.

AUTHOR CONTRIBUTIONS

AB prepared the draft text of the manuscript and figures. UZ and AD developed the concept and edited the manuscript and figures. AB, AD, and UZ finalized the specific parts. FG developed a scheme for single-cell tools describing. FG, FC, and SM participated in discussing the concept of the article, advised writing sections of the manuscript and design of figures. All authors contributed to the manuscript and approved the submitted version.

FUNDING

The manuscript concept and analytical review of literature were supported by the Russian Foundation for Basic Research (Project No. 20-04-01112). A NoSelf-UNINA grant project financially supported AB elaborating the general framework for modeling plant morphogenesis. The access to the database of single-cell datasets and its overall analysis was performed using resources of Shared Computational Facilities Center Bioinformatics supported by the State Budget Program (Project No. 0259-2021-0009).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.652974/full#supplementary-material

REFERENCES

Ali, S., Khan, N., and Xie, L. (2020). Molecular and hormonal regulation of leaf morphogenesis in Arabidopsis. *Int. J. Mol. Sci.* 21:5132. doi: 10.3390/ijms21145132

Amezquita, R. A., Lun, A. T., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., et al. (2020). Orchestrating single-cell analysis with bioconductor. *Nat. Methods* 17, 137–145. doi: 10.1038/s41592-019-0654-x

Banwarth-Kuhn, M., Nematbakhsh, A., Rodriguez, K. W., Snipes, S., Rasmussen, C. G., Reddy, G. V., et al. (2019). Cell-based model of the generation and maintenance of the shape and structure of the multilayered shoot apical meristem of *Arabidopsis thaliana*. *Bull. Math. Biol.* 81, 3245–3281. doi: 10.1007/s11538-018-00547-z

Barrio, R. A., Romero-Arias, J. R., Noguez, M. A., Azpeitia, E., Ortiz-Gutiérrez, E., Hernández-Hernández, V., et al. (2013). Cell patterns emerge from coupled chemical and physical fields with cell proliferation dynamics: the Arabidopsis thaliana root as a study system. *PLoS Comput. Biol.* 9:e1003026. doi: 10.1371/journal.pcbi.1003026

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314

Beltrán, J., Wamboldt, Y., Sanchez, R., LaBrant, E. W., Kundariya, H., Virdi, K. S., et al. (2018). Specialized plastids trigger tissue-specific signaling for systemic stress response in plants. *Plant Physiol.* 178, 672–683. doi: 10.1104/pp.18.00804

Bezrutczyk, M., Zöllner, N. R., Kruse, C. P., Hartwig, T., Lautwein, T., Köhrer, K., et al. (2021). Evidence for phloem loading via the abaxial bundle sheath cells in maize leaves. *Plant Cell*. koaa055. doi: 10.1093/plcell/koaa055

Bidhendi, A. J., and Geitmann, A. (2016). Relating the mechanics of the primary plant cell wall to morphogenesis. *J. Exp. Bot.* 67, 449–461. doi: 10.1093/jxb/erv535

Bidhendi, A. J., and Geitmann, A. (2019). Methods to quantify primary plant cell wall mechanics. *J. Exp. Bot.* 70, 3615–3648. doi: 10.1093/jxb/erz281

Biswas, S., and Barma, S. (2020). A large-scale optical microscopy image dataset of potato tuber for deep learning based plant cell assessment. *Sci. Data* 7, 1–11. doi: 10.1038/s41597-020-00706-9

Blinov, M. L., Schaff, J. C., Vasilescu, D., Moraru, I. I., Bloom, J. E., and Loew, L. M. (2017). Compartmental and spatial rule-based modeling with virtual cell. *Biophys. J.* 113, 1365–1372. doi: 10.1016/j.bpj.2017.08.022

Bobrovskikh, A., Zubairova, U., Kolodkin, A., and Doroshkov, A. (2020). Subcellular compartmentalization of the plant antioxidant system: an integrated overview. *PeerJ* 8:e9451. doi: 10.7717/peerj.9451

Bowman, J. (2012). *Arabidopsis: An Atlas of Morphology and Development*. New York, NY: Springer Science & Business Media.

Campbell, K. R., and Yau, C. (2017). Probabilistic modeling of bifurcations in single-cell gene expression data using a bayesian mixture of factor analyzers. *Wellcome Open Res.* 2:19. doi: 10.12688/wellcomeopenres.11087.1

Chan, T. E., Stumpf, M. P., and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* 5, 251–267. doi: 10.1016/j.cels.2017.08.014

Cowan, A. E., Moraru, I. I., Schaff, J. C., Slepchenko, B. M., and Loew, L. M. (2012). "Spatial modeling of cell signaling networks," in *Methods in Cell Biology*, Vol. 110, eds A. R. Asthagiri, and A. P. Arkin (Elsevier; Academic Press), 195–221. doi: 10.1016/B978-0-12-388403-9.00008-4

Crawford, K. M., and Zambryski, P. C. (2001). Non-targeted and targeted protein movement through plasmodesmata in leaves in different developmental and physiological states. *Plant Physiol.* 125, 1802–1812. doi: 10.1104/pp.125.4.1802

De Rybel, B., Adibi, M., Breda, A. S., Wendrich, J. R., Smit, M. E., Novák, O., et al. (2014). Integration of growth and patterning during vascular tissue formation in Arabidopsis. *Science* 345:6197. doi: 10.1126/science.1255215

de Souza, L. P., Borghi, M., and Fernie, A. (2020). Plant single-cell metabolomics-challenges and perspectives. *Int. J. Mol. Sci.* 21(23):8987. doi: 10.3390/ijms21238987

De Vos, D., Dzhurakhalov, A., Draelants, D., Bogaerts, I., Kalve, S., Prinsen, E., et al. (2012). Towards mechanistic models of plant organ growth. *J. Exp. Bot.* 63, 3325–3337. doi: 10.1093/jxb/ers037

De Vos, D., Vissenberg, K., Broeckhove, J., and Beemster, G. T. (2014). Putting theory to the test: which regulatory mechanisms can drive realistic growth of a root? *PLoS Comput. Biol.* 10:e1003910. doi: 10.1371/journal.pcbi.1003910

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., and Timmermans, M. C. (2019). Spatiotemporal developmental trajectories in the Arabidopsis root revealed using high-throughput single-cell RNA sequencing. *Dev. cell* 48, 840–852. doi: 10.1016/j.devcel.2019.02.022

Dolan, L., Janmaat, K., Willemsen, V., Linstead, P., Poethig, S., Roberts, K., et al. (1993). Cellular organisation of the *Arabidopsis thaliana* root. *Development* 119, 71–84. doi: 10.1242/dev.119.1.71

Du, F., and Jiao, Y. (2020). Mechanical control of plant morphogenesis: concepts and progress. *Curr. Opin. Plant Biol.* 57, 16–23. doi: 10.1016/j.pbi.2020.05.008

Dupuy, L., Mackenzie, J., and Haseloff, J. (2010). Coordination of plant cell division and expansion in a simple morphogenetic system. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2711–2716. doi: 10.1073/pnas.0906322107

Dupuy, L., Mackenzie, J., Rudge, T., and Haseloff, J. (2008). A system for modelling cell-cell interactions during plant morphogenesis. *Ann. Bot.* 101, 1255–1265. doi: 10.1093/aob/mcm235

Durbak, A., Yao, H., and McSteen, P. (2012). Hormone signaling in plant development. *Curr. Opin. Plant Biol.* 15, 92–96. doi: 10.1016/j.pbi.2011.12.004

Erguvan, Ö., Louveaux, M., Hamant, O., and Verger, S. (2019). Imagej surfcut: a user-friendly pipeline for high-throughput extraction of cell contours from 3d image stacks. *BMC Biol.* 17:38. doi: 10.1186/s12915-019-0657-1

Farmer, A., Thibivilliers, S., Ryu, K. H., Schiefelbein, J., and Libault, M. (2021). Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level. *Molecular Plant*. 14, 372–383. doi: 10.1016/j.molp.2021.01.001

Fitzgibbon, J., Beck, M., Zhou, J., Faulkner, C., Robatzek, S., and Oparka, K. (2013). A developmental framework for complex plasmodesmata formation revealed by large-scale imaging of the Arabidopsis leaf epidermis. *Plant Cell* 25, 57–70. doi: 10.1105/tpc.112.105890

Fitzgibbon, J., Bell, K., King, E., and Oparka, K. (2010). Super-resolution imaging of plasmodesmata using three-dimensional structured illumination microscopy. *Plant Physiol.* 153, 1453–1463. doi: 10.1104/pp.110.157941

Formosa-Jordan, P., Teles, J., and Jönsson, H. (2018). "Single-cell approaches for understanding morphogenesis using computational morphodynamics," in *Mathematical Modelling in Plant Biology*, ed R. Morris (Cham: Springer International Publishing), 87–106. doi: 10.1007/978-3-319-99070-5_6

Fricker, M. D. (2016). Quantitative redox imaging software. *Antioxid. Redox Signal.* 24, 752–762. doi: 10.1089/ars.2015.6390

Gajdanowicz, P., Michard, E., Sandmann, M., Rocha, M., Corrêa, L. G. G., Ramirez-Aguilar, S. J., et al. (2011). Potassium (k+) gradients serve as a mobile energy source in plant vascular tissues. *Proc. Natl. Acad. Sci. U.S.A.* 108, 864–869. doi: 10.1073/pnas.1009777108

Gierlinger, N., Keplinger, T., and Harrington, M. (2012). Imaging of plant cell walls by confocal raman microscopy. *Nat. Protoc.* 7, 1694–1708. doi: 10.1038/nprot.2012.092

Gilmore, I. S., Heiles, S., and Pieterse, C. L. (2019). Metabolic imaging at the single-cell scale: recent advances in mass spectrometry imaging. *Annu. Rev. Anal. Chem.* 12, 201–224. doi: 10.1146/annurev-anchem-061318-115516

Goh, T. (2019). Long-term live-cell imaging approaches to study lateral root formation in *Arabidopsis thaliana*. *Microscopy* 68, 4–12. doi: 10.1093/jmicro/dfy135

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13:845. doi: 10.1038/nmeth.3971

Heinlein, M., and Epel, B. L. (2004). Macromolecular transport and signaling through plasmodesmata. *Int. Rev. Cytol.* 235, 93–164. doi: 10.1016/S0074-7696(04)35003-5

Hester, S. D., Belmonte, J. M., Gens, J. S., Clendenon, S. G., and Glazier, J. A. (2011). A multi-cell, multi-scale model of vertebrate segmentation and somite formation. *PLoS Comput. Biol.* 7:e1002155. doi: 10.1371/journal.pcbi.1002155

Hong, L., Dumond, M., Zhu, M., Tsugawa, S., Li, C.-B., Boudaoud, A., et al. (2018). Heterogeneity and robustness in plant morphogenesis: from cells to organs. *Annu. Rev. Plant Biol.* 69, 469–495. doi: 10.1146/annurev-arplant-042817-040517

Hossain, M. S., Joshi, T., and Stacey, G. (2015). System approaches to study root hairs as a single cell plant model: current status and future perspectives. *Front. Plant Sci.* 6:363. doi: 10.3389/fpls.2015.00363

Huang, K., Batish, M., Teng, C., Harkess, A., Meyers, B. C., and Caplan, J. L. (2020). "Quantitative fluorescence *in situ* hybridization detection of plant mrnas with single-molecule resolution," in *RNA Tagging*, ed M. Heinlein (New York, NY: Springer), 23–33. doi: 10.1007/978-1-0716-0712-1_2

Interian, R., Rodriguez-Ramos, R., Valdés-Ravelo, F., Ramirez-Torres, A., Ribeiro, C., and Conci, A. (2017). Tumor growth modelling by cellular automata. *Math. Mech. Complex Syst.* 5, 239–259. doi: 10.2140/memocs.2017.5.239

Irrthum, A., Wehenkel, L., Geurts, P., et al. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* 5:e12776. doi: 10.1371/journal.pone.0012776

Jackson, M. D., Xu, H., Duran-Nebreda, S., Stamm, P., and Bassel, G. W. (2017). Topological analysis of multicellular complexity in the plant hypocotyl. *eLife* 6:e26023. doi: 10.7554/eLife.26023

Jean-Baptiste, K., McFaline-Figueroa, J. L., Alexandre, C. M., Dorrity, M. W., Saunders, L., Bubb, K. L., et al. (2019). Dynamics of gene expression in single root cells of *Arabidopsis thaliana*. *Plant Cell* 31, 993–1011. doi: 10.1105/tpc.18.00785

Jeannin-Girardon, A., Ballet, P., and Rodin, V. (2015). Large scale tissue morphogenesis simulation on heterogenous systems based on a flexible biomechanical cell model. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12, 1021–1033. doi: 10.1109/TCBB.2015.2418994

Kerstens, M., Strauss, S., Smith, R., and Willemsen, V. (2020). "From stained plant tissues to quantitative cell segmentation analysis with morphographx," in *Plant Embryogenesis*, ed M. Bayer (New York, NY: Springer), 63–83. doi: 10.1007/978-1-0716-0342-0_6

Kim, J.-Y., Symeonidi, E., Pang, T. Y., Denyer, T., Weidauer, D., Bezrutczyk, M., et al. (2021). Distinct identities of leaf phloem cells revealed by single cell transcriptomics. *Plant Cell*. koaa060:1–34. doi: 10.1093/plcell/koaa060

Kim, J. T., Jakobsen, S., Natarajan, K. N., and Won, K.-J. (2020). Tenet: gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data. *Nucleic Acids Res.* 49, 1–11. doi: 10.1093/nar/gkaa1014

Kiss, A., Moreau, T., Mirabet, V., Calugaru, C. I., Boudaoud, A., and Das, P. (2017). Segmentation of 3D images of plant tissues at multiple scales using the level set method. *Plant Methods* 13, 1–11. doi: 10.1186/s13007-017-0264-5

Kováčik, J., and Babula, P. (2017). Fluorescence microscopy as a tool for visualization of metal-induced oxidative stress in plants. *Acta Physiol. Plant.* 39:157. doi: 10.1007/s11738-017-2455-0

Kragler, F. (2013). Plasmodesmata: intercellular tunnels facilitating transport of macromolecules in plants. *Cell Tissue Res.* 352, 49–58. doi: 10.1007/s00441-012-1550-1

Kremer, A., Lippens, S., Bartunkova, S., Asselbergh, B., Blanpain, C., Fendrych, M., et al. (2015). Developing 3D SEM in a broad biological context. *J. Microsc.* 259, 80–96. doi: 10.1111/jmi.12211

Krzesłowska, M., Timmers, A. C., Mleczek, M., Niedzielski, P., Rabeda, I., Woźny, A., et al. (2019). Alterations of root architecture and cell wall modifications in tilia cordata miller (linden) growing on mining sludge. *Environ. Pollut.* 248, 247–259. doi: 10.1016/j.envpol.2019.02.019

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* 21, 1–35. doi: 10.1186/s13059-020-1926-6

Lavrekha, V. V., Pasternak, T., Palme, K., and Mironova, V. V. (2020). "3D analysis of mitosis distribution pattern in the plant root tip with irocs toolbox," in *Plant Stem Cells*, eds M. Naseem, T. Dandekar (New York, NY: Springer), 119–125. doi: 10.1007/978-1-0716-0183-9_13

Legland, D., Arganda-Carreras, I., and Andrey, P. (2016). Morpholibj: integrated library and plugins for mathematical morphology with imagej. *Bioinformatics* 32, 3532–3534. doi: 10.1093/bioinformatics/btw413

Lev-Yadun, S. (2015). Plant development: Cell movement relative to each other is both common and very important. *Plant Signal. Behav.* 10:e991566. doi: 10.4161/15592324.2014.991566

Li, L., Zhang, Q., and Huang, D. (2014). A review of imaging techniques for plant phenotyping. *Sensors* 14, 20078–20111. doi: 10.3390/s141120078

Libault, M., Pingault, L., Zogli, P., and Schiefelbein, J. (2017). Plant systems biology at the single-cell level. *Trends Plant Sci.* 22, 949–960. doi: 10.1016/j.tplants.2017.08.006

Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., and Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* 16, 243–245. doi: 10.1038/s41592-018-0308-4

Liu, Q., Liang, Z., Feng, D., Jiang, S., Wang, Y., Du, Z., et al. (2021). Transcriptional landscape of rice roots at the single-cell resolution. *Mol. Plant* 14, 384–394. doi: 10.1016/j.molp.2020.12.014

Locato, V., and De Gara, L. (2018). "Programmed cell death in plants: an overview," in *Plant Programmed Cell Death*, eds L. De Gara, V. Locato (New York, NY: Springer), 1–8. doi: 10.1007/978-1-4939-7668-3_1

Lockhart, J. A. (1965). An analysis of irreversible plant cell elongation. *J. Theor. Biol.* 8, 264–275. doi: 10.1016/0022-5193(65)90077-9

Lopez-Anido, C. B., Vatén, A., Smoot, N. K., Sharma, N., Guo, V., Gong, Y., et al. (2021). Single-cell resolution of lineage trajectories in the Arabidopsis stomatal lineage and developing leaf. *Dev Cell.* 56, 1043–1055. doi: 10.1016/j.devcel.2021.03.014

Lucas, W. J., and Lee, J.-Y. (2004). Plasmodesmata as a supracellular control network in plants. *Nat. Rev. Mol. Cell Biol.* 5, 712–726. doi: 10.1038/nrm1470

Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Mol. Syst. Biol.* 15:e8746. doi: 10.15252/msb.20188746

Luo, K.-R., Huang, N.-C., and Yu, T.-S. (2018). Selective targeting of mobile mrnas to plasmodesmata for cell-to-cell movement. *Plant Physiol.* 177, 604–614. doi: 10.1104/pp.18.00107

Ma, A., McDermaid, A., Xu, J., Chang, Y., and Ma, Q. (2020). Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol.* 38, 1007–1022. doi: 10.1016/j.tibtech.2020.02.013

Madhikar, P., Åström, J., Westerholm, J., and Karttunen, M. (2018). *CellSim3D*: Gpu accelerated software for simulations of cellular growth and division in three dimensions. *Comput. Phys. Commun.* 232, 206–213. doi: 10.1016/j.cpc.2018.05.024

Mai, D., Dürr, J., Palme, K., and Ronneberger, O. (2014). "Accurate detection in volumetric images using elastic registration based validation," in *German Conference on Pattern Recognition*, eds X. Jiang, J. Hornegger, and R. Koch (Cham: Springer International Publishing), 453–463. doi: 10.1007/978-3-319-11752-2_37

McFaline-Figueroa, J. L., Trapnell, C., and Cuperus, J. T. (2020). The promise of single-cell genomics in plants. *Curr. Opin. Plant Biol.* 54, 114–121. doi: 10.1016/j.pbi.2020.04.002

Merks, R. M., Guravage, M., Inzé, D., and Beemster, G. T. (2011). Virtualleaf: an open-source framework for cell-based modeling of plant tissue growth and development. *Plant Physiol.* 155, 656–666. doi: 10.1104/pp.110.167619

Mironova, V., Omelyanchuk, N., Novoselova, E., Doroshkov, A., Kazantsev, F., Kochetov, A., et al. (2012). Combined *in silico/in vivo* analysis of mechanisms providing for root apical meristem self-organization and maintenance. *Ann. Bot.* 110, 349–360. doi: 10.1093/aob/mcs069

Mjolsness, E. (2019). Prospects for declarative mathematical modeling of complex biological systems. *Bull. Math. Biol.* 81, 3385–3420. doi: 10.1007/s11538-019-00628-7

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., et al. (2019). Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 35, 2159–2161. doi: 10.1093/bioinformatics/bty916

Morales-Tapia, A., and Cruz-Ramírez, A. (2016). Computational modeling of auxin: a foundation for plant engineering. *Front. Plant Sci.* 7:1881. doi: 10.3389/fpls.2016.01881

Moraru, I. I., Schaff, J. C., Slepchenko, B. M., Blinov, M., Morgan, F., Lakshminarayana, A., et al. (2008). Virtual cell modelling and simulation software environment. *IET Syst. Biol.* 2, 352–362. doi: 10.1049/iet-syb:20080102

Muraro, D., Mellor, N., Pound, M. P., Lucas, M., Chopard, J., Byrne, H. M., et al. (2014). Integration of hormonal signaling networks and mobile micrornas is required for vascular patterning in Arabidopsis roots. *Proc. Natl. Acad. Sci. U.S.A.* 111, 857–862. doi: 10.1073/pnas.1221766111

Nagai, T., and Honda, H. (2001). A dynamic cell model for the formation of epithelial tissues. *Philos. Mag. B* 81, 699–719. doi: 10.1080/13642810108205772

Nelson, C. M., Jean, R. P., Tan, J. L., Liu, W. F., Sniadecki, N. J., Spector, A. A., et al. (2005). Emergent patterns of growth controlled by multicellular form and mechanics. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11594–11599. doi: 10.1073/pnas.0502575102

Nguyen, L. H., and Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLoS Comput. Biol.* 15:e1006907. doi: 10.1371/journal.pcbi.1006907

Nicolas, W. J., Grison, M. S., Trépout, S., Gaston, A., Fouché, M., Cordeliéres, F. P., et al. (2017). Architecture and permeability of post-cytokinesis plasmodesmata lacking cytoplasmic sleeves. *Nat. Plants* 3:17082. doi: 10.1038/nplants.2017.82

Nikolaev, S., Zubairova, U., Penenko, A., Mjolsness, E., Shapiro, B., and Kolchanov, N. (2013). Model of structuring the stem cell niche in shoot apical meristem of Arabidopsis thaliana. *Doklady Biol. Sci.* 452, 316–319. doi: 10.1134/S0012496613050104

Omari, M. K., Lee, J., Faqeerzada, M. A., Joshi, R., Park, E., and Cho, B.-K. (2020). Digital image-based plant phenotyping: a review. *Korean J. Agric. Sci.* 47, 119–130. doi: 10.34133/2020/4152816

Omelyanchuk, N., Kovrizhnykh, V., Oshchepkova, E., Pasternak, T., Palme, K., and Mironova, V. (2016). A detailed expression map of the pin1 auxin transporter in Arabidopsis thaliana root. *BMC Plant Biol.* 16:5. doi: 10.1186/s12870-015-0685-0

Onal, S., Turker-Burhan, M., Bati-Ayaz, G., Yanik, H., and Pesen-Okvur, D. (2020). Breast cancer cells and macrophages in a paracrine-juxtacrine loop. *Biomaterials* 267:120412. doi: 10.1016/j.biomaterials.2020.120412

Ortega, J. K. (2010). Plant cell growth in tissue. *Plant Physiol.* 154, 1244–1253. doi: 10.1104/pp.110.162644

Osborne, J. M., Fletcher, A. G., Pitt-Francis, J. M., Maini, P. K., and Gavaghan, D. J. (2017). Comparing individual-based approaches to modelling the self-organization of multicellular tissues. *PLoS Comput. Biol.* 13:e1005387. doi: 10.1371/journal.pcbi.1005387

Palm, M. M., and Merks, R. M. (2015). "Large-scale parameter studies of cell-based models of tissue morphogenesis using compucell3d or virtualleaf," in *Tissue Morphogenesis*, ed C. Nelson (New York, NY: Springer), 301–322. doi: 10.1007/978-1-4939-1164-6_20

Pasternak, T., Haser, T., Falk, T., Ronneberger, O., Palme, K., and Otten, L. (2017). A 3d digital atlas of the *Nicotiana tabacum* root tip and its use to investigate changes in the root apical meristem induced by the Agrobacterium 6b oncogene. *Plant J.* 92, 31–42. doi: 10.1111/tpj.13631

Petegrosso, R., Li, Z., and Kuang, R. (2020). Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief. Bioinform.* 21, 1209–1223. doi: 10.1093/bib/bbz063

Pfluger, J., and Zambryski, P. C. (2001). Cell growth: the power of symplastic isolation. *Curr. Biol.* 11, R436–R439. doi: 10.1016/S0960-9822(01)00254-8

Pradal, C., Dufour-Kowalski, S., Boudon, F., Fournier, C., and Godin, C. (2008). Openalea: a visual programming and component-based software platform for plant modelling. *Funct. Plant Biol.* 35, 751–760. doi: 10.1071/FP08084

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17, 147–154. doi: 10.1038/s41592-019-0690-6

Retta, M. A., Abera, M. K., Berghuijs, H. N., Verboven, P., Struik, P. C., and Nicolaï, B. M. (2020). *In silico* study of the role of cell growth factors in photosynthesis using a virtual leaf tissue generator coupled to a microscale photosynthesis gas exchange model. *J. Exp. Bot.* 71, 997–1009. doi: 10.1093/jxb/erz451

Rhee, S. Y., Birnbaum, K. D., and Ehrhardt, D. W. (2019). Towards building a plant cell atlas. *Trends Plant Sci.* 24, 303–310. doi: 10.1016/j.tplants.2019.01.006

Rich-Griffin, C., Stechemesser, A., Finch, J., Lucas, E., Ott, S., and Schäfer, P. (2020). Single-cell transcriptomics: a high-resolution avenue for plant functional genomics. *Trends Plant Sci.* 25, 186–197. doi: 10.1016/j.tplants.2019.10.008

Roeder, A. H., Cunha, A., Ohno, C. K., and Meyerowitz, E. M. (2012). Cell cycle regulates cell type in the Arabidopsis sepal. *Development* 139, 4416–4427. doi: 10.1242/dev.082925

Romero-Arias, J. R., Hernández-Hernández, V., Benitez, M., Alvarez-Buylla, E. R., and Barrio, R. A. (2017). Model of polar auxin transport coupled to mechanical forces retrieves robust morphogenesis along the Arabidopsis root. *Phys. Rev. E* 95:032410. doi: 10.1103/PhysRevE.95.032410

Rounds, C. M., and Bezanilla, M. (2013). Growth mechanisms in tip-growing plant cells. *Annu. Rev. Plant Biol.* 64, 243–265. doi: 10.1146/annurev-arplant-050312-120150

Rudge, T. J., Steiner, P. J., Phillips, A., and Haseloff, J. (2012). Computational modeling of synthetic microbial biofilms. *ACS Synthet. Biol.* 1, 345–352. doi: 10.1021/sb300031n

Ryu, K. H., Huang, L., Kang, H. M., and Schiefelbein, J. (2019). Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiol*. 179, 1444–1456. doi: 10.1104/pp.18.01482

Sablowski, R. (2016). Coordination of plant cell growth and division: collective control or mutual agreement? *Curr. Opin. Plant Biol*. 34, 54–60. doi: 10.1016/j.pbi.2016.09.004

Saini, S., Sharma, I., Kaur, N., and Pati, P. K. (2013). Auxin: a master regulator in plant root development. *Plant Cell Rep*. 32, 741–757. doi: 10.1007/s00299-013-1430-5

Sassi, M., Ali, O., Boudon, F., Cloarec, G., Abad, U., Cellier, C., et al. (2014). An auxin-mediated shift toward growth isotropy promotes organ formation at the shoot meristem in Arabidopsis. *Curr. Biol*. 24, 2335–2342. doi: 10.1016/j.cub.2014.08.036

Satterlee, J. W., Strable, J., and Scanlon, M. J. (2020). Plant stem-cell organization and differentiation at single-cell resolution. *Proc. Natl. Acad. Sci. U.S.A*. 117, 33689–33699. doi: 10.1073/pnas.2018788117

Sauer, M., and Friml, J. (2010). "Immunolocalization of proteins in plants," in *Plant Developmental Biology*, eds L. Hennig, and C. Köhler (Totowa, NJ: Springer; Humana Press), 253–263. doi: 10.1007/978-1-60761-765-5_17

Schmidt, T., Pasternak, T., Liu, K., Blein, T., Aubry-Hivet, D., Dovzhenko, A., et al. (2014). The IROCS toolbox-3D analysis of the plant root apical meristem at cellular resolution. *Plant J*. 77, 806–814. doi: 10.1111/tpj.12429

Seerangan, K., van Spoordonk, R., Sampathkumar, A., and Eng, R. C. (2020). Long-term live-cell imaging techniques for visualizing pavement cell morphogenesis. *Methods Cell Biol*. 160, 365–380. doi: 10.1016/bs.mcb.2020.04.007

Sevilem, I., Yadav, S. R., and Helariutta, Y. (2015). "Plasmodesmata: channels for intercellular signaling during plant growth and development," in *Plasmodesmata*, ed M. Heinlein (New York, NY: Springer), 3–24. doi: 10.1007/978-1-4939-1523-1_1

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13, 2498–2504. doi: 10.1101/gr.1239303

Shapiro, B. E., Levchenko, A., Meyerowitz, E. M., Wold, B. J., and Mjolsness, E. D. (2003). Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations. *Bioinformatics* 19, 677–678. doi: 10.1093/bioinformatics/btg042

Shapiro, B. E., Meyerowitz, E., and Mjolsness, E. (2013). Using cellzilla for plant growth simulations at the cellular level. *Front. Plant Sci*. 4:408. doi: 10.3389/fpls.2013.00408

Shapiro, B. E., Tobin, C., Mjolsness, E., and Meyerowitz, E. M. (2015). Analysis of cell division patterns in the Arabidopsis shoot apical meristem. *Proc. Natl. Acad. Sci. U.S.A*. 112, 4815–4820. doi: 10.1073/pnas.1502588112

Shaw, R., Tian, X., and Xu, J. (2020). Single-cell transcriptome analysis in plants: advances and challenges. *Mol. Plant*. 14, 115–126. doi: 10.1016/j.molp.2020.10.012

She, W., Baroux, C., and Grossniklaus, U. (2018). "Cell-type specific chromatin analysis in whole-mount plant tissues by immunostaining," in *Plant Chromatin Dynamics*, ed M. Bemer, and C. Baroux (New York, NY: Springer), 443–454. doi: 10.1007/978-1-4939-7318-7_25

Shulse, C. N., Cole, B. J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., et al. (2019). High-throughput single-cell transcriptome profiling of plant cell types. *Cell Rep*. 27, 2241–2247. doi: 10.1016/j.celrep.2019.04.054

Smolarkiewicz, M., and Dhonukshe, P. (2013). Formative cell divisions: principal determinants of plant morphogenesis. *Plant Cell Physiol*. 54, 333–342. doi: 10.1093/pcp/pcs175

Stopka, A., Kokic, M., and Iber, D. (2019). Cell-based simulations of biased epithelial lung growth. *Phys. Biol*. 17:016006. doi: 10.1088/1478-3975/ab5613

Swat, M. H., Thomas, G. L., Belmonte, J. M., Shirinifard, A., Hmeljak, D., and Glazier, J. A. (2012). "Multi-scale modeling of tissues using compucell3d," in *Methods in Cell Biology*, eds A. R. Asthagiri and A. P. Arkin (Elsevier), 325–366. doi: 10.1016/B978-0-12-388403-9.00013-8

Swat, M. H., Thomas, G. L., Shirinifard, A., Clendenon, S. G., and Glazier, J. A. (2015). Emergent stratification in solid tumors selects for reduced cohesion of tumor cells: a multi-cell, virtual-tissue model of tumor evolution using compucell3d. *PLoS ONE* 10:e0127972. doi: 10.1371/journal.pone.0127972

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 47, D607–D613. doi: 10.1093/nar/gky1131

Tanaka, S., Sichau, D., and Iber, D. (2015). Lbibcell: a cell-based simulation environment for morphogenetic problems. *Bioinformatics* 31, 2340–2347. doi: 10.1093/bioinformatics/btv147

Tauriello, G., Meyer, H. M., Smith, R. S., Koumoutsakos, P., and Roeder, A. H. (2015). Variability and constancy in cellular growth of Arabidopsis sepals. *Plant Physiol*. 169, 2342–2358. doi: 10.1104/pp.15.00839

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res*. 25, 1491–1498. doi: 10.1101/gr.190595.115

Turco, G. M., Rodriguez-Medina, J., Siebert, S., Han, D., Valderrama-Gómez, M., Á., et al. (2019). Molecular mechanisms driving switch behavior in xylem cell differentiation. *Cell Rep*. 28, 342–351. doi: 10.1016/j.celrep.2019.06.041

Uyttewaal, M., Burian, A., Alim, K., Landrein, B., Borowska-Wykret, D., Dedieu, A., et al. (2012). Mechanical stress acts via katanin to amplify differences in growth rate between adjacent cells in Arabidopsis. *Cell* 149, 439–451. doi: 10.1016/j.cell.2012.02.048

Vaahtera, L., Schulz, J., and Hamann, T. (2019). Cell wall integrity maintenance during plant development and interaction with the environment. *Nat. Plants* 5, 924–932. doi: 10.1038/s41477-019-0502-0

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., et al. (2020). A scalable scenic workflow for single-cell gene regulatory network analysis. *Nat. Protoc*. 15, 2247–2276. doi: 10.1038/s41596-020-0336-2

Van Liedekerke, P., Palm, M., Jagiella, N., and Drasdo, D. (2015). Simulating tissue mechanics with agent-based models: concepts, perspectives and some novel results. *Comput. Part. Mech*. 2, 401–444. doi: 10.1007/s40571-015-0082-3

van Mourik, S., Kaufmann, K., van Dijk, A. D., Angenent, G. C., Merks, R. M., and Molenaar, J. (2012). Simulation of organ patterning on the floral meristem using a polar auxin transport model. *PLoS ONE* 7:e28762. doi: 10.1371/journal.pone.0028762

Vijayan, A., Tofanelli, R., Strauss, S., Cerrone, L., Wolny, A., Strohmeier, J., et al. (2021). A digital 3D reference atlas reveals cellular growth patterns shaping the *Arabidopsis ovule*. *eLife*. 10:e63262. doi: 10.7554/eLife.63262

Vincenot, C. E., Giannino, F., Rietkerk, M., Moriya, K., and Mazzoleni, S. (2011). Theoretical considerations on the combined use of system dynamics and individual-based modeling in ecology. *Ecol. Modell*. 222, 210–218. doi: 10.1016/j.ecolmodel.2010.09.029

von Wangenheim, D., Fangerau, J., Schmitz, A., Smith, R. S., Leitte, H., Stelzer, E. H., et al. (2016). Rules and self-organizing properties of post-embryonic plant organ cell division patterns. *Curr. Biol*. 26, 439–449. doi: 10.1016/j.cub.2015.12.047

Voß, U., Bishopp, A., Farcot, E., and Bennett, M. J. (2014). Modelling hormonal response and development. *Trends Plant Sci*. 19, 311–319. doi: 10.1016/j.tplants.2014.02.004

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14, 414–416. doi: 10.1038/nmeth.4207

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 38(Suppl_2), W214–W220. doi: 10.1093/nar/gkq537

Warner, C. A., Biedrzycki, M. L., Jacobs, S. S., Wisser, R. J., Caplan, J. L., and Sherrier, D. J. (2014). An optical clearing technique for plant tissues allowing deep imaging and compatible with fluorescence microscopy. *Plant Physiol*. 166, 1684–1687. doi: 10.1104/pp.114.244673

Wiese, R., Eiglsperger, M., and Kaufmann, M. (2004). "*yFiles*– visualization and automatic layout of graphs," in *Graph Drawing Software*, eds M. Jünger, and P. Mutzel (Berlin; Heidelberg: Springer), 173–191. doi: 10.1007/978-3-642-18638-7_8

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol*. 19:15. doi: 10.1186/s13059-017-1382-0

Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A. V., Louveaux, M., et al. (2020). Accurate and versatile 3D segmentation of plant tissues at cellular resolution. *eLife*. 9:e57613. doi: 10.7554/eLife.57613.sa2

Wuyts, N., Palauqui, J.-C., Conejero, G., Verdeil, J.-L., Granier, C., and Massonnet, C. (2010). High-contrast three-dimensional imaging of the Arabidopsis leaf

enables the analysis of cell dimensions in the epidermis and mesophyll. *Plant Methods* 6, 1–14. doi: 10.1186/1746-4811-6-17

Xu, X., Crow, M., Rice, B. R., Li, F., Harris, B., Liu, L., et al. (2021). Single-cell rna sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery. *Dev. Cell.* 56:557–568.E6. doi: 10.1016/j.devcel.2020.12.015

Xuan, W., Band, L. R., Kumpf, R. P., Van Damme, D., Parizot, B., De Rop, G., et al. (2016). Cyclic programmed cell death stimulates hormone signaling and root development in Arabidopsis. *Science* 351, 384–387. doi: 10.1126/science.aad2776

Yadav, S. R., Yan, D., Sevilem, I., and Helariutta, Y. (2014). Plasmodesmata-mediated intercellular signaling during plant growth and development. *Front. Plant Sci.* 5:44. doi: 10.3389/fpls.2014.00044

Yamamoto, K., Takahashi, K., Caputi, L., Mizuno, H., Rodriguez-Lopez, C. E., Iwasaki, T., et al. (2019). The complexity of intercellular localisation of alkaloids revealed by single-cell metabolomics. *N. Phytol.* 224, 848–859. doi: 10.1111/nph.16138

Yarbrough, J. M., Himmel, M. E., and Ding, S.-Y. (2009). Plant cell wall characterization using scanning probe microscopy techniques. *Biotechnol. Biofuels* 2, 1–11. doi: 10.1186/1754-6834-2-17

Zeise, I., Heiner, Z., Holz, S., Joester, M., Büttner, C., and Kneipp, J. (2018). Raman imaging of plant cell walls in sections of cucumis sativus. *Plants* 7:7. doi: 10.3390/plants7010007

Zhang, L., McEvoy, D., Le, Y., and Ambrose, C. (2020). Live imaging of microtubule organization, cell expansion, and intercellular space formation in Arabidopsis leaf spongy mesophyll cells. *Plant Cell.* 50:koaa036. doi: 10.1093/plcell/koaa036

Zhang, T., Zheng, Y., and Cosgrove, D. J. (2016). Spatial organization of cellulose microfibrils and matrix polysaccharides in primary plant cell walls as imaged by multichannel atomic force microscopy. *Plant J.* 85, 179–192. doi: 10.1111/tpj.13102

Zhang, T.-Q., Xu, Z.-G., Shang, G.-D., and Wang, J.-W. (2019). A single-cell rna sequencing profiles the developmental landscape of Arabidopsis root. *Mol. Plant* 12, 648–660. doi: 10.1016/j.molp.2019.04.004

Zubairova, U., Nikolaev, S., Penenko, A., Podkolodnyy, N., Golushko, S., Afonnikov, D., et al. (2016). Mechanical behavior of cells within a cell-based model of wheat leaf growth. *Front. Plant Sci.* 7:1878. doi: 10.3389/fpls.2016.01878

Zubairova, U. S., Verman, P. Y., Oshchepkova, P. A., Elsukova, A. S., and Doroshkov, A. V. (2019). LSM-W2: laser scanning microscopy worker for wheat leaf surface morphology. *BMC Syst. Biol.* 13:22. doi: 10.1186/s12918-019-0689-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 15 September 2021
doi: 10.3389/fgene.2021.721229

Dice-XMBD: Deep Learning-Based Cell Segmentation for Imaging Mass Cytometry

Xu Xiao [1,2†], Ying Qiao [1†], Yudi Jiao [1], Na Fu [1], Wenxian Yang [3], Liansheng Wang [1], Rongshan Yu [1,2,3*] and Jiahuai Han [2,4*]*

[1] Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China, [2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China, [3] Aginome Scientific, Xiamen, China, [4] School of Medicine, Xiamen University, Xiamen, China

Highly multiplexed imaging technology is a powerful tool to facilitate understanding the composition and interactions of cells in tumor microenvironments at subcellular resolution, which is crucial for both basic research and clinical applications. Imaging mass cytometry (IMC), a multiplex imaging method recently introduced, can measure up to 100 markers simultaneously in one tissue section by using a high-resolution laser with a mass cytometer. However, due to its high resolution and large number of channels, how to process and interpret the image data from IMC remains a key challenge to its further applications. Accurate and reliable single cell segmentation is the first and a critical step to process IMC image data. Unfortunately, existing segmentation pipelines either produce inaccurate cell segmentation results or require manual annotation, which is very time consuming. Here, we developed Dice-XMBD[1], a Deep learnIng-based Cell sEgmentation algorithm for tissue multiplexed imaging data. In comparison with other state-of-the-art cell segmentation methods currently used for IMC images, Dice-XMBD generates more accurate single cell masks efficiently on IMC images produced with different nuclear, membrane, and cytoplasm markers. All codes and datasets are available at https://github.com/xmuyulab/Dice-XMBD.

Keywords: imaging mass cytometry, multiplexed imaging, single cell segmentation, U-net, knowledge distillation, digital pathology

OPEN ACCESS

Edited by:
Min Wu,
*Institute for Infocomm Research (A*STAR), Singapore*

Reviewed by:
Mengwei Li,
*Singapore Immunology Network (A*STAR), Singapore*
Yuan Zhu,
China University of Geosciences Wuhan, China

***Correspondence:**
Liansheng Wang
lswang@xmu.edu.cn
Rongshan Yu
rsyu@xmu.edu.cn
Jiahuai Han
jhan@xmu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *06 June 2021*
Accepted: *30 July 2021*
Published: *15 September 2021*

Citation:
Xiao X, Qiao Y, Jiao Y, Fu N, Yang W, Wang L, Yu R and Han J (2021) Dice-XMBD: Deep Learning-Based Cell Segmentation for Imaging Mass Cytometry. Front. Genet. 12:721229. doi: 10.3389/fgene.2021.721229

1. INTRODUCTION

Analysis of the heterogeneity of cells is critical to discover the complexity and factuality of life system. Recently, single-cell sequencing technologies have been increasingly used in the research of developmental physiology and disease (Stubbington et al., 2017; Papalexi and Satija, 2018; Potter, 2018; Lähnemann et al., 2020), but the spatial context of individual cells in the tissue is lost due to tissue dissociation in these technologies. On the other hand, traditional immunohistochemistry (IHC) and immunofluorescence (IF) preserve spatial context but the number of biomarkers is limited. The development of multiplex IHC/IF (mIHC/mIF) technologies has enabled the simultaneous detection of multiple biomarkers and preserves spatial information, such as cyclic IHC/IF and metal-based multiplex imaging technologies (Zrazhevskiy and Gao, 2013; Angelo et al., 2014; Giesen et al., 2014; Tan et al., 2020). Imaging mass cytometry (IMC) (Giesen et al., 2014; Chang et al., 2017), one of

[1]XMBD: Xiamen Big Data, a biomedical open software initiative in the National Institute for Data Science in Health and Medicine, Xiamen University, China.

metal-based mIHC technologies, uses a high-resolution laser with a mass cytometer and makes the measurement of 100 markers possible.

IMC has been utilized in studies of cancer and autoimmune disorders (Giesen et al., 2014; Damond et al., 2019; Ramaglia et al., 2019; Wang et al., 2019; Böttcher et al., 2020). Due to its high resolution and large number of concurrent marker channels available, IMC has been proven to be highly effective in identifying the complex cell phenotypes and interactions coupled with spatial locations. Thus, it has become a powerful tool to study tumor microenvironments and discover the underlying disease-relevant mechanisms (Brähler et al., 2018; Ali et al., 2020; Aoki et al., 2020; de Vries et al., 2020; Dey et al., 2020; Jackson et al., 2020; Zhang et al., 2020; Schwabenland et al., 2021). Apart from using IMC techniques alone, several other technologies, such as RNA detection *in situ* and 3D imaging, have been combined with IMC to expand its applicability and utility (Schulz et al., 2018; Bouzekri et al., 2019; Catena et al., 2020; Flint et al., 2020).

The IMC data analysis pipeline typically starts with single cell segmentation followed by tissue/cell type identification (Carpenter et al., 2006; Sommer et al., 2011; Liu et al., 2019). As the first step of an IMC data processing pipeline, the accuracy of single cell segmentation plays a significant role in determining the quality and the reliability of the biological results from an IMC study. Existing IMC cell segmentation methods include both unsupervised and supervised algorithms. Unsupervised cell segmentation, such as the watershed algorithm implemented in CellProfiler (Carpenter et al., 2006), does not require user inputs for model training. However, the segmentation results are not precise in particular when cells are packed closely or they are in complicated shapes. To achieve better segmentation results, supervised methods use a set of images annotated with pixel-level cell masks to train a segmentation classifier. However, the manual annotation task is very time consuming and expensive as well since it is normally done by pathologists or experienced staff with necessary knowledge in cell annotation. Particularly, for multiplexing cellular imaging methods such as IMC, their channel configurations including the total number of markers and markers selection are typically study dependent. Therefore, manual annotation may need to be performed repeatedly for each study to adapt the segmentation model to different IMC channel configurations, which can be impractical.

To overcome this limitation, a hybrid workflow combining unsupervised and supervised learning methods for cell segmentation was proposed (Ali et al., 2020). This hybrid workflow uses Ilastik (Sommer et al., 2011), an interactive image processing tool, to generate a probability map based on multiple rounds of user inputs and adjustments. In each round, a user only needs to perform a limited number of annotations on regions where the probability map generated based on previous annotations is not satisfactory. CellProfiler is then used to perform the single cell segmentation based on the probability map once the result from Ilastik is acceptable. This hybrid workflow significantly reduces manual annotation workload and has gained popularity in many recent IMC studies (Damond et al., 2019; Böttcher et al., 2020; de Vries et al., 2020; Jackson et al., 2020; Schwabenland et al., 2021). However, the annotation process still needs to be performed by experienced staff repeatedly for each IMC study, which is very inconvenient. In addition, the reproducibility of the experimental results obtained from this approach can be an issue due to the per-study, interactive training process used in creating the single cell masks. Hence, a more efficient, fully automated single cell segmentation method for IMC data without compromising the segmentation accuracy is necessary for IMC to gain broader applications in biomedical studies.

Convolutional neural networks (CNNs) have been successfully used for natural image segmentation and recently applied in biomedical image applications (Shen et al., 2017; Zhang et al., 2018; Andrade et al., 2019; Vicar et al., 2019). CNN-based U-Net was developed for pixel-wise cell segmentation of mammalian cells (Ronneberger et al., 2015). It has been demonstrated that the U-Net architecture and its variants such as Unet++ (Zhou et al., 2018), 3D Unet (Çiçek et al., 2016), and V-Net (Milletari et al., 2016) can obtain high segmentation accuracy. Motivated by the good performance of U-Nets in cell segmentation (Van Valen et al., 2016; Hollandi et al., 2020; Salem et al., 2020), we developed Dice-XMBD, a deep neural network (DNN)-based cell segmentation method for multichannel IMC images. Dice-XMBD is marker agnostic and can perform cell segmentation for IMC images of different channel configurations without modification. To achieve this goal, Dice-XMBD first merges multiple-channel IMC images into two channels, namely, a nuclear channel containing proteins originated from cell nucleus, and a cell channel containing proteins originated from cytoplasm and cell membrane. Channels of proteins with ambiguous locations are ignored by Dice-XMBD for segmentation as they contribute little to the segmentation results. Furthermore, to mitigate the annotation workload, we adopted the knowledge distillation learning framework (Hinton et al., 2015) in training Dice-XMBD, where the training labels were generated using Ilastik with interactive manual annotations as a teacher model. We used four IMC datasets of different channel configurations to evaluate the performance of Dice-XMBD and the results show that it can generate highly accurate cell segmentation results that are comparable to those from manual annotation for IMC images from both the same and different datasets to the training dataset, validating its applicability for generic IMC image segmentation tasks.

2. MATERIALS AND METHODS

2.1. Overview of the Pipeline

In Dice-XMBD, we used a U-Net-based pixel classification model to classify individual pixels of an IMC image to their cellular origins, namely, nuclei, cytoplasm/membrane, or background. The classification model outputs pixel-level probability values for each class, which were then input to CellProfiler (version 3.1.0) to produce the final cell segmentation masks (**Figure 1**).

The ground truth cell segmentation of IMC images is in general not available. To obtain the training labels, we generated pixel probability maps using an iterative manual annotation process with Ilastik on the training IMC dataset. Furthermore, the same iterative manual annotation process was performed on the testing IMC datasets to produce the ground

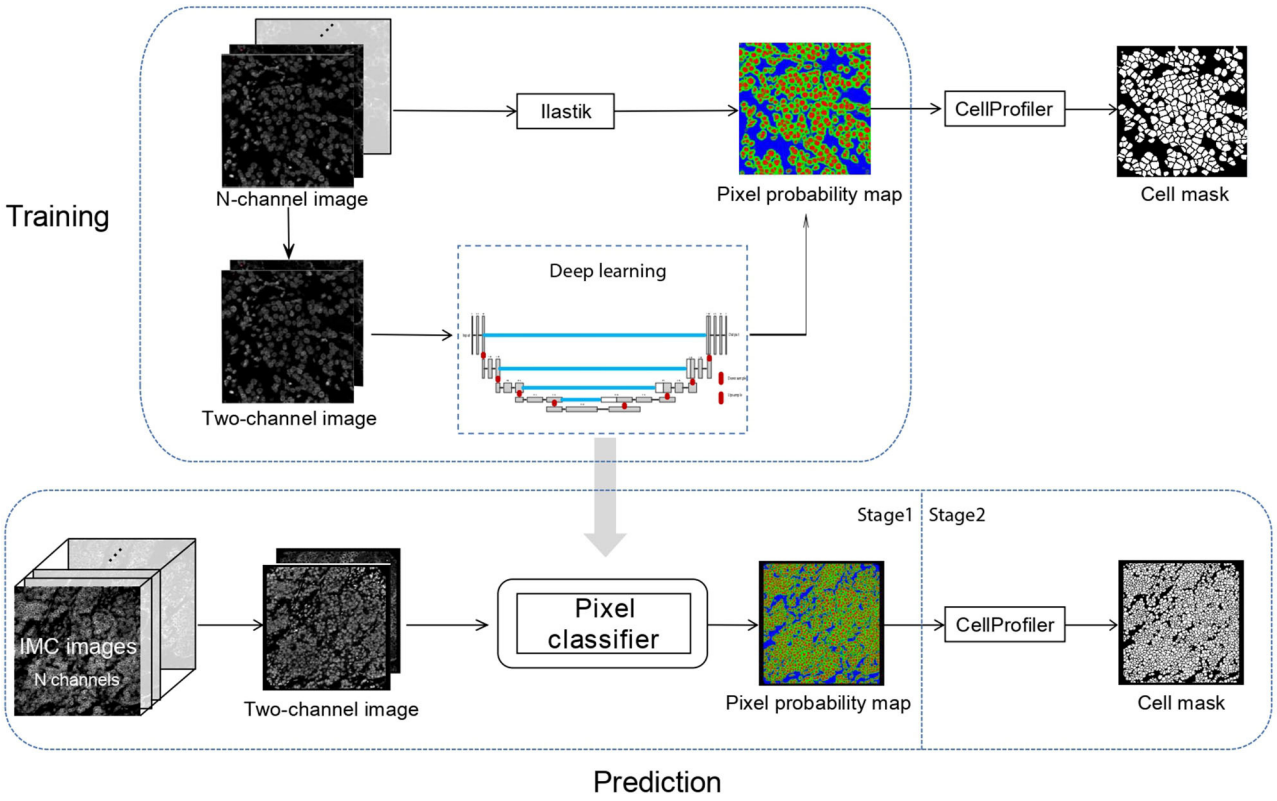


FIGURE 1 | Dice-XMBD workflow. Imaging mass cytometry (IMC) images are combined into 2-channel images containing nuclear and membrane/cytoplasm proteins expression information. In stage 1, the pixel probability maps of the input 2-channel images are predicted using a semi-supervised learning model based on U-Net architecture. The training data were generated from Ilastik by an iterative interactive annotation process. In stage 2, the cell segmentation masks are generated from the pixel probability maps using the propagation method in CellProfiler.

truth pixel probability maps, which were used by CellProfiler to produce the ground truth cell segmentation masks for performance evaluation.

Note that to obtain a generic pixel classifier that can be used across IMC datasets of different channel configurations, channels of different proteins were combined based on their cellular origins into two channels, namely, nuclear and cell (membrane/cytoplasmic) channels. Channels of proteins without specific cellular locations were ignored by Dice-XMBD. The pixel classification model was trained using the combined two-channel images as input. Likewise, the same preprocessing was used at the prediction stage to produce the two-channel (nuclear/cell) images as input to the pixel classification model. Of note, although the prediction may be performed on images with different markers, the channels were always combined based on their origins so that pixel classification was performed based on the two channels of putative protein locations rather than channels of individual proteins.

2.2. Training and Evaluation Datasets

We used four IMC image datasets in this study. BRCA1 and BRCA2 (Ali et al., 2020) contain 548 and 746 images from patients with breast cancer with 36 and 33 markers, respectively. T1D1 (Damond et al., 2019) and T1D2 (Wang et al., 2019) contain 839 and 754 images from patients with type I diabetes with 34 markers. Dice-XMBD was trained on a subset of BRCA1 dataset ($n = 348$) with 200 held-out images reserved for validation and testing. To test the generalization ability of Dice-XMBD, we also tested the trained model on the other three independent IMC datasets (BRCA2, T1D1, and T1D2).

2.3. Generating Ground Truth Cell Masks

The ground truth pixel probability maps and the cell masks used for model training and evaluation were generated using Ilastik and CellProfiler. We used the smallest brush size (1 pixel) in annotating the image to avoid annotating a group of neighboring pixels of different classes. To mitigate the manual workload, the annotation was performed in an interactive manner, where the random forest prediction model of Ilastik was updated regularly during annotation to produce an uncertainty map indicating the confidence level of the classification results produced by the prediction model. The annotation was then guided by the uncertainty map to focus on the regions with high uncertainty iteratively, until the overall uncertainty values were low except for regions of which the boundaries were visually indistinguishable.

The initial annotation was performed on a randomly selected subset of the dataset. After the initial annotation, we loaded all the images from the dataset into Ilastik to calculate their

uncertainty maps, and then selected those with the highest average uncertainty values for further annotation. This process was iterated until the uncertainty values of all images converged, that is, the average uncertainty value over all images did not decrease significantly for three consecutive iterations.

In the end, we annotated 49 images in BRCA1 to train the model in Ilastik. We then imported all the images of the BRCA1 dataset into Ilastik for batch processing and export their corresponding pixel classification probability maps for training Dice-XMBD. The probability maps were further input to CellProfiler to produce the ground truth cell segmentation. In CellProfiler, we used the "IdentifyPrimaryObjects" module to segment the cell nuclei and used the "IdentifySecondaryObjects" to segment the cell membranes using the propagation method. The output masks from CellProfiler are regarded as ground truth cell segmentation of the dataset for performance evaluation.

We also generated the ground truth cell masks of the other three datasets by the same iterative procedure separately for testing the generalization ability of Dice-XMBD. During the process, 72 images in BRCA2, 39 images in T1D1, and 67 images in T1D2 were manually annotated.

2.4. Training the U-Net Cell Segmentation Model

2.4.1. Image Preprocessing

The multiplexed IMC images were first merged into two channels by averaging the per-pixel values from the selected membrane and nuclear channels. After merging channels, the input IMC images were then preprocessed by hot pixel removal, dynamic range conversion, normalization, and image cropping/padding into fix-sized patches. First, we applied a 5 × 5 low-pass filter on the image to remove hot pixels. If the difference between an image pixel value and the corresponding filtered value was larger than a preset threshold (50 in our experiments), the pixel would be regarded as a hot pixel and its value would be replaced by the filtered value. As the dynamic range of pixels values differs among IMC images of different batches and different channels, we further min-max normalized all images to [0,255] to remove such batch effect as:

$$x_{ij}^{'} = \frac{x_{ij} - X_{min}}{X_{max} - X_{min}} * 255, \quad (1)$$

where x_{ij} denotes the pixel value in one channel, and X_{max} and X_{min} denote the maximum and minimum values in the channel. Of note, as the pixel values in IMC images have a high dynamic range, transforming the pixel values from its dynamic range to [0, 255] would suffer from detail suppression by one or few extremely large values. Therefore, we thresholded the image pixel values at 99.7% percentile for each image before normalization.

Finally, we merged all the nuclear channels into one consolidated nuclear channel, and membrane/cytoplasmic channels into one cell channel, by averaging on all channel images with pre-selected sets of protein markers, respectively. We converted the merged two-channel images into patches of 512 × 512 pixels. Image boundary patches that are smaller than the target patch size are padded to target size. For the padded pixels, we set the pixel values of both channels to 0 and the pixel type as background.

2.4.2. Data Augmentation

Data augmentation is an effective strategy to reduce overfitting and enhance the robustness of the trained models, especially when training data are insufficient. We applied the following data augmentation methods on the input images before feeding to our U-Net-based pixel classification network.

First, photometric transformations including contrast stretching and intensity adjustments were used. For contrast stretching, we changed the level of contrast by multiplication with a factor randomly drawn from the range of [0.5, 1.5]. Similarly, for intensity adjustments we changed the level of intensities by multiplication with a factor randomly drawn from the range of [0.5, 1.5]. Geometric transformations including image flipping and rotation were used. For flipping, we implemented random horizontal or vertical flipping. For rotation, the rotating angle is randomly distributed in the range of [−180, 180]. Note that geometric transformations were applied to pairs of input and output images of the network. We also injected random Gaussian noise to the two input channels of the input images. Examples of data augmentation are shown in **Supplementary Figures 1**, **2**.

2.4.3. Constructing a Pixel Classification Model

The U-Net pixel classification network is an end-to-end fully convolutional network and contains two paths. The contracting path (or the encoder) uses a typical CNN architecture. Each block in the contracting path consists of two successive 3 × 3 convolution layers followed by a Rectified Linear Unit (ReLU) activation and a 2 × 2 max-pooling layer. This block is repeated four times. In the symmetric expansive path (or the decoder), at each stage the feature map is upsampled using 2 × 2 up-convolution. To enable precise localization, the feature map from the corresponding layer in the contracting path is cropped and concatenated onto the upsampled feature map, followed by two successive 3 × 3 convolutions and ReLU activation. At the final stage, an additional 1 × 1 convolution is applied to reduce the feature map to the required number of output channels. Three output channels are used in our case for nuclei, membrane, and background, respectively. As we output the probability map, the values are converted into the range of [0, 1] using the Sigmoid function.

2.4.4. Loss Function

We take the binary cross-entropy (BCE) as the loss function, which is defined as:

$$loss(y, \hat{y}) = -\frac{1}{N} \sum_{i=0}^{N} (y_i * log(\hat{y}_i) + (1 - y_i) * log(1 - \hat{y}_i)), \quad (2)$$

where N represents the total number of pixels in an image, y_i denotes the ground truth pixel probability, and $\hat{y}_i$ denotes the predicted pixel probability. The cross-entropy loss compares the predicted probabilities with the ground truth values. The loss is minimized during the training process.

2.5. Model Evaluation

In a binary cell mask, "1" represents cell boundary and "0" denotes cell interior or exterior. For every pixel in an image, true positive (TP) and true negative (TN) mean that the predicted pixel classification is the same as its label in the labeled (i.e., the ground truth) mask, while false positive (FP) and false negative (FN) mean that a pixel is misclassified. To evaluate pixel-level accuracy, we calculated the number of TP pixels and FP pixels based on the predicted and labeled binary masks.

We further evaluated model performance at the cell level. We calculated the intersection over union (IOU) on cells from predicted and labeled cell masks to determine if they are the same cell, and then counted the TP and FP cells. First, we filtered out all cells with IOU below 0.1 from the predicted cells. These cells are identified as FPs. The other cells from the predicted cell mask could be either TP or FP. If a predicted cell only overlaps with one true cell (i.e., a cell from the labeled cell mask), we assume that the cell is segmented accurately (TP). If a true cell cannot find a predicted cell, the "missing" cell is denoted as FN. When multiple predicted cells are assigned to the same true cell, we consider this as a split error. If multiple true cells are matched to the same predicted cell, we consider those predicted cells as merge errors. For simplicity, split errors and merge errors are counted as FPs. Four standard indices are measured as follows:

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{5}$$

$$Jaccard = \frac{TP}{TP + FP + FN}. \tag{6}$$

To investigate the effect of different segmentation methods on downstream analysis, an unsupervised clustering method (Phenograph Levine et al., 2015, Python package, v1.5.7) was applied to the high-dimensional single cell expression data processed from each different method under comparison, and the labeled ground truth cell mask, separately. Prior to clustering, single cell protein expressions were quantified by the mean pixel values, and then these values were censored at 99th percentile and transformed with arcsinh function. Scaled high-dimensional single cells were clustered into several groups based on selected markers as from the original publication of each individual dataset. Based on the expressions of cell-specific markers, the cell types of the clusters were identified among T cells (CD3), CD4 T cell (CD4), CD8 T cell (CD8a), B cell (CD20), macrophage (CD68), endothelial cell (CD31), and so on. By comparing the cell annotation from different segmentation methods (predicted cell mask) and the labeled cell mask, the cell annotation accuracy was calculated as n_{same}/N_{total}. Here, n_{same} is the number of correctly predicted cells, which are cells that correctly overlapped with the corresponding cells in the labeled mask (i.e., TP cells), and annotated to the same cell types, N_{total} is the total number of cells from the predicted mask.

3. RESULTS

3.1. Dice-XMBD Enables Automatic Cell Segmentation

We trained our U-Net cell segmentation model using the BRCA1 dataset with 348 images as the training set and 100 images as the validation set. A complete held out test set with 100 images was used to test model performance within one dataset. We further applied the trained model directly on the other three IMC image datasets to evaluate the cross-dataset performance of the model. For performance evaluation, we computed standard indices (Recall, Precision, F1-score, and Jaccard index) for both pixel-level and cell-level accuracies (see section 2).

We compared Dice-XMBD with a generic whole-cell segmentation method across six imaging platforms, Mesmer (Greenwald et al., 2021), which used a deep learning-based algorithm trained on a large, annotated image dataset to segment single cells and nuclei separately. A trained Mesmer model was tested with combined nuclear and cell channels, which is the same as the input to Dice-XMBD. Meanwhile, we compared with three commonly used segmentation methods implemented in CellProfiler with default parameters: distance, watershed, and propagation. These methods first locate nuclei as primary objects, and then the membrane proteins are added together into an image as input to recognize cells. The distance method does not use any membrane proteins information and simply defines cell membrane by expanding several pixels around nuclei. The watershed method computes intensity gradients on the Sobel transformed image to identify boundaries between cells (Vincent and Soille, 1991), while the propagation method defines cell boundaries by combining the distance of the nearest primary object and the intensity gradients of cell membrane image (Jones et al., 2005). Hereafter, we refer to these three CellProfile-based methods as CP_distance, CP_watershed, and CP_propagation, respectively.

Results show that Dice-XMBD outperformed all other benchmarked methods with highest accuracy on pixel level (F1 score = 0.92, Jaccard index = 0.85) (**Figure 2A**). We also observed that CP_distance obtained the highest recall (Recall = 0.95) but lowest precision (Precision = 0.66), which means that it can identify almost every pixel correctly in the labeled mask but only 66% of predicted pixels were accurate.

In terms of cell-level performance, we first counted cells per image from predicted and labeled cell masks. The prediction result from Dice-XMBD showed highest correlation with the ground truth (Pearson correlation = 0.998) among all methods tested. Mesmer (Pearson correlation = 0.955) and CellProfiler (Pearson correlation = 0.981) also achieved high correlation with the ground truth. However, Mesmer tended to predict less cells while CellProfiler was more likely to over-split cells, as shown in **Figures 2B,C**. Moreover, **Figure 2C** shows that Dice-XMBD had the best prediction performance (F1-score = 0.867) considering precision (Precision = 0.856, percent of cells that were correctly predicted) and recall (Recall = 0.880, percent of true cells that are predicted) than Mesmer (F1-score = 0.557) and CellProfiler (F1-score = 0.567, 0.563, and 0.561 for CP_distance, CP_watershed, and CP_propagation, respectively). We further checked the IOU

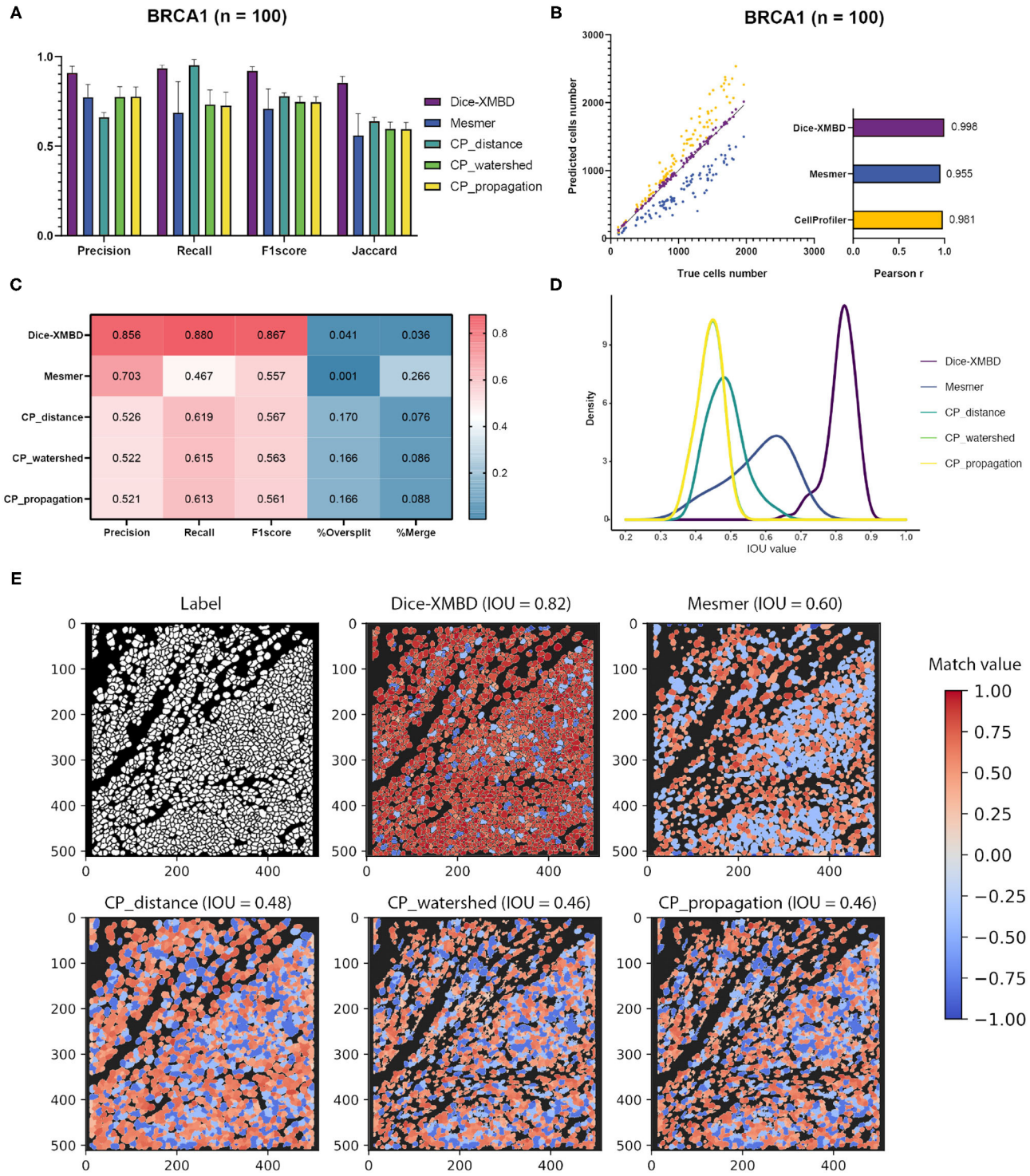


FIGURE 2 | Dice-XMBD enables automatic single cell segmentation. **(A)** Pixel prediction performance comparison of Dice-XMBD, Mesmer, and CellProfiler (CP_distance, CP_watershed, CP_propagation). All data in bar plots are presented as mean ±SD. **(B)** Pearson correlations between the number of predicted cells and labeled cells per image. Note that the number of cells predicted from three CellProfiler methods are the same (here denoted as CellProfiler). **(C)** Cell prediction performance comparison. %Oversplit and %Merge denote the percentage of oversplits and merge errors in predictions. **(D)** Density plots showing the distribution of mean IOU values of matched cells per image. Note that the plots for CP_watershed and CP_propagation overlapped. **(E)** An example of labeled and predicted single cell masks from benchmarked methods. The title of each subfigure shows the method and the mean IOU value of all matched cell pairs in the predicted mask with regard to the labeled cell mask. Match value represents the IOU value for one-to-one cell pairs identified in the labeled and predicted cell masks. Note that computed IOU values are in the range of [0,1]. To better visualize FP cells, we use –0.4 and –0.8 to represent merged cells (multiple true cells matched to one predicted cell) and split cells (multiple predicted cells matched to one true cell), and –1 to represent all other FP cells in the predicted mask.

distribution of all one-to-one cell pairs (predicted and true cells), **Figure 2D** demonstrates that most matched cell pairs predicted from Dice-XMBD were highly overlapping (mean = 0.815, median = 0.821), followed by Mesmer where most matched pairs are only half area of overlap (mean = 0.579, median = 0.595). An example of BRCA1 shown in **Figure 2E** demonstrates that Dice-XMBD prediction was far superior to other benchmarked methods since it contained most cells with high matched values.

3.2. Dice-XMBD Enables Generic IMC Image Segmentation

The key idea of this study was to generate an IMC-specific single cell segmentation model across different datasets with multiple proteins. We selected three independent IMC datasets generated from different labs to test the generalization ability of Dice-XMBD. Apart from the benchmarked methods mentioned above, we also included the Ilastik model trained from BRCA1 annotations in our comparison. **Figure 3A** shows that Dice-XMBD outperformed all the other methods, followed by Ilastik. Moreover, the performance of cells prediction from Dice-XMBD was the best and the most stable for all three datasets, while Ilastik and Mesmer tended to under-predict cells. CellProfiler predicted less cells in BRCA2 and over-predicted cells in two T1D datasets, as shown in **Figures 3B,C**. Furthermore, Dice-XMBD predictions contained most of the cells with IOU value higher than 0.8 (**Figure 3D** and **Supplementary Figure 3**).

3.3. Dice-XMBD Enables Accurate Downstream Biological Analysis

To investigate the influence of segmentation accuracy on downstream analysis, we clustered single cells resulting from different segmentation methods separately using Phenograph and compared the clustering results. Taking the result from single cells obtained from Dice-XMBD segmentation on BRCA1 dataset as an example, these cells can be clustered into 26 distinct clusters [**Figure 4A**, t-distributed stochastic neighbor embedding (t-SNE) visualization in **Figure 4B**]. Based on the scaled mean expression for each cluster, we were able to annotate Cluster 3 as T cells, Cluster 18 as B cells, Cluster 16 as macrophage, and the remaining clusters to other cell types which may include tumor cells, stromal cells, or endothelial cells (**Figure 4C**). We performed the same clustering and annotation process on single cells obtained from other segmentation methods and the ground truth segmentation on all three datasets separately as well. For two T1D datasets [T1D1 (Damond et al., 2019) and T1D2 (Wang et al., 2019)], CD4 T cells, CD8 T cells, and CD31+ endothelial cells were identified based on their selected markers.

We compared the concordance of cell fractions based on annotations from different segmentation methods (prediction) versus those from ground truth segmentation (ground truth) (**Figure 4D** and **Supplementary Figures 4A–7A**). On BRCA1 dataset, Dice-XMBD performed better compared with all other segmentation methods on overall results and results of certain cell types (**Figure 4D**). Significantly, two CellProfiler-based methods (CP_watershed, $R^2 = 0.85$ and CP_propagation, $R^2 = 0.85$) showed inferior performance in reproducing cell fraction results in macrophage while Dice-XMBD still achieved an $R^2 = 0.99$ in this cell types. CP_distance delivered reasonable performance in macrophage, but was still inferior to Dice-XMBD on T cell. Similar results can be observed on other datasets as well. For example, for the T1D1 dataset, CD4 T cells were poorly predicted by Ilastik ($R^2 = 0.043$) and CP_distance ($R^2 = 0.055$) (**Supplementary Figure 6A**). For the T1D2 dataset, endothelial cells were poorly predicted by Ilastik ($R^2 = 0.58$) and macrophage cells were poorly predicted by Mesmer ($R^2 = 0.033$). On the other hand, Dice-XMBD delivered highly consistent prediction results across all cell types in all datasets except for T cell in BRCA2 dataset, where all methods did not perform well.

In addition to cell fraction, we also evaluated the annotation accuracy of individual cells for each method (**Figure 4E** and **Supplementary Figures 4B–7B**), which is important for spatially related analysis of single cell data such as neighborhood analysis. Dice-XMBD achieved the highest cell annotation accuracies among all segmentation methods on overall results (**Figure 4E**), and performed as well as or better than other methods on all individual cell types in all datasets (**Supplementary Figures 4B–7B**).

3.4. Generalization Ability of Dice-XMBD

To investigate the impact of the training data on the segmentation performance of Dice-XMBD, we trained Dice-XMBD using different training datasets, and evaluated the performance of the resulting models on other IMC datasets used in this study. Results show that segmentation performance in terms of pixel-level accuracy were in fact very similar among these models (**Supplementary Tables 1–4**). We further asked if the performance of Dice-XMBD could be improved by training on multiple datasets. Interestingly, the model did not consistently perform better when more than one datasets were combined as the training set (**Supplementary Tables 1–4**). All together, these results suggest that by using location specific channels, Dice-XMBD were highly robust to different training datasets, and a Dice-XMBD model trained on one dataset can be well generalized to segmentation tasks on other IMC datasets.

Of note, in our approach, the channels of same locations were simply averaged without applying any weighting scheme to produce the location specific channels. We tried to min-max-normalize the selected channels before averaging so that all selected channels contributed equally to the combined channels. However, the pixel-level accuracy dropped on all datasets, albeit at different levels of degradation on different datasets (**Supplementary Tables 1–4**). As different channels may contain different levels of information to the final segmentation results, combining them with equal weights may not be the optimal approach. However, how to find the optimal weighting combination of different channels remains an open question that deserves further exploration.

4. DISCUSSION

Highly multiplexed single cell imaging technologies such as IMC are becoming increasingly important tools for both basic biomedical and clinical research. These tools can unveil complex

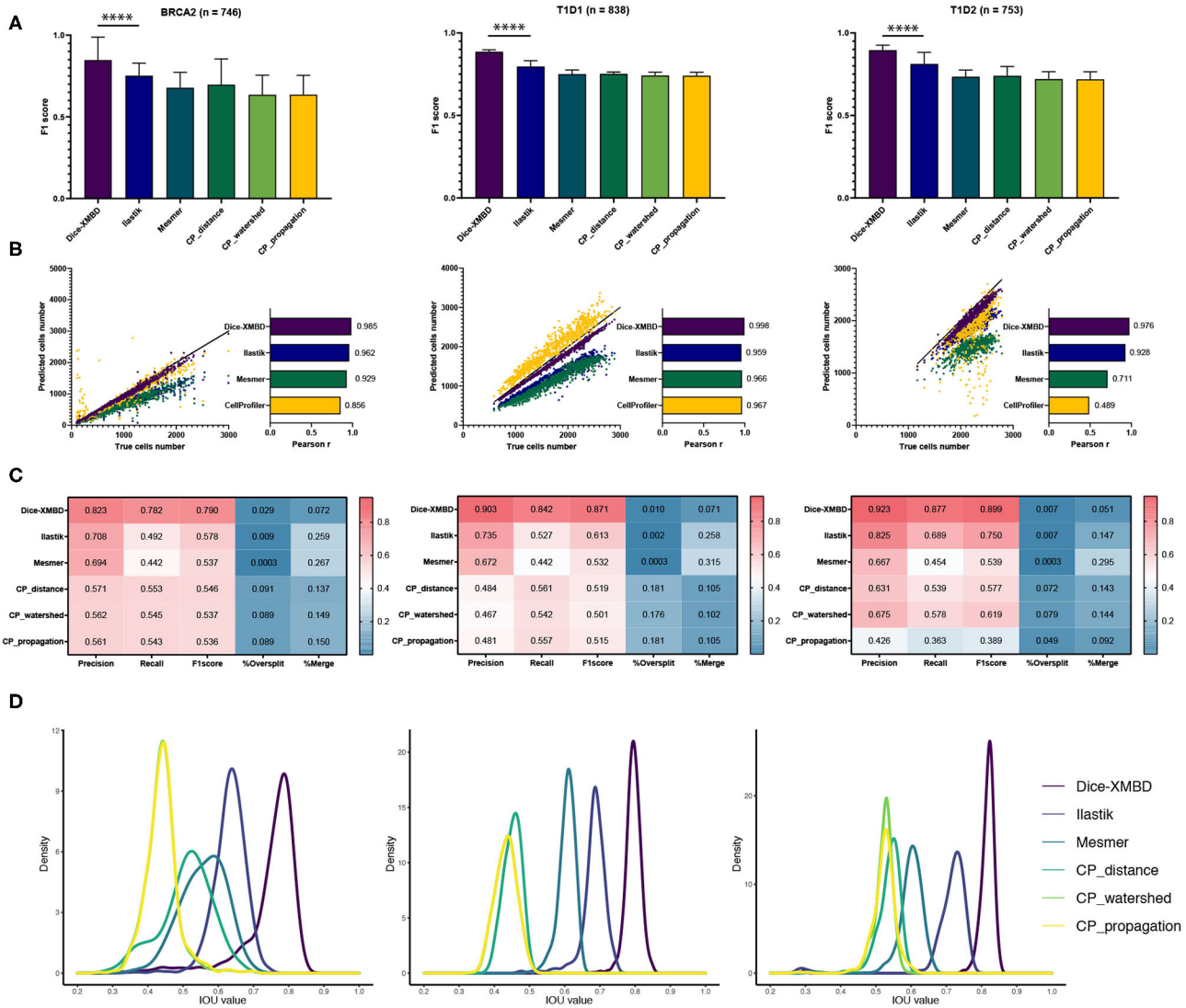


FIGURE 3 | Dice-XMBD enables generic imaging mass cytometry (IMC) image segmentation. Left: BCRA2, middle: T1D1, right: T1D2. **(A)** Pixel prediction performance comparison of Dice-XMBD, Ilastik, Mesmer, and CellProfiler (CP_distance, CP_watershed, CP_propagation). All data in bar plots are presented as mean ±SD. **(B)** Pearson correlations between the number of predicted cells and labeled cells per image. Note that the number of cells predicted from three CellProfiler methods are the same (here denoted as CellProfiler). **(C)** Heatmaps of cells prediction performance of six benchmarked methods. %Oversplit and %Merge denote the percentage of oversplits and merge errors in predictions. **(D)** Density plots showing the distribution of mean IOU values of matched cells per image. Note that the plots for CP_watershed and CP_propagation overlapped for BRCA2 and T1D1.

single-cell phenotypes and their spatial context at unprecedented details, providing a solid base for further exploration in cancer, diabetes, and other complex diseases. Nevertheless, cell segmentation has become a major bottleneck in analyzing multiplexed images. Conventional approaches rely on intensities of protein markers to identify different cellular structures such as nuclei, cytoplasm, and membrane. Unfortunately, the intensity values of these markers are strongly cell type-specific and may vary from cells to cells. In addition, the staining also shows variability across images or datasets. As a result, the accuracy and robustness of the segmentation results are far from optimal. On the other hand, high-order visual features including spatial distribution of markers, textures, and gradients are relevant to visually identify subcellular structures by human. However, these features are not considered in conventional methods to improve the cell segmentation results.

The DNN-based image segmentation approaches provide an opportunity to leverage high-order visual features at cellular level for better segmentation results. Unfortunately, they require a significant amount of annotation data that are in general difficult to acquire. In addition, the highly variable channel configurations of multiplexed images impose another important obstacle to

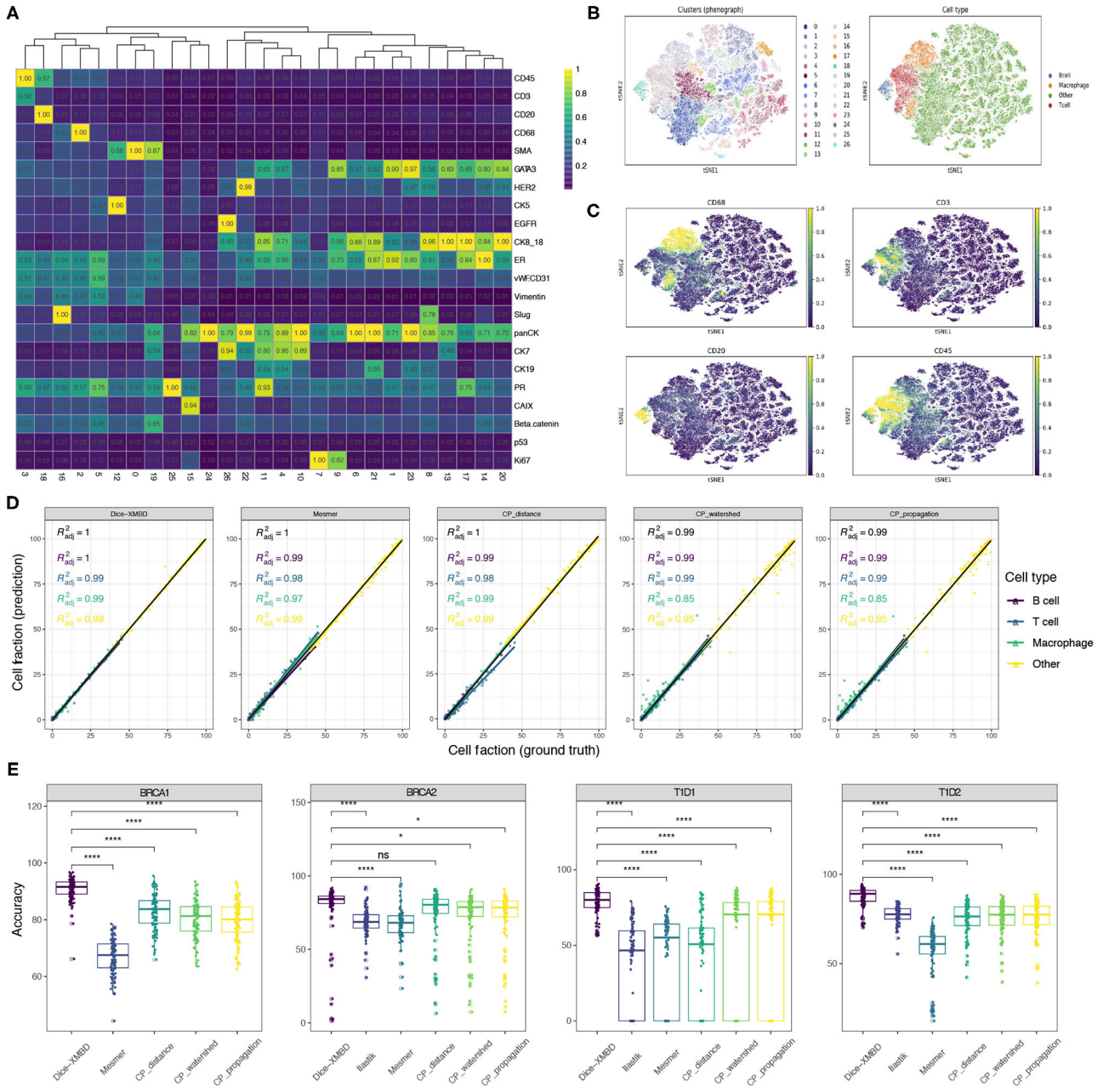


FIGURE 4 | Dice-XMBD enables accurate downstream biological analysis. **(A)** Heatmap showing median values of normalized markers expression in each Phenograph cluster. **(B)** tSNE map representing high-dimensional single cells colored by Phenograph clusters (left) and cell types (right). **(C)** tSNE map representing single cells colored by cell-type-specific markers expression (CD68 for macrophage, CD45 and CD3 for T cells, CD45 and CD20 for B cells). Single cells on **(A–C)** were from BRCA1 dataset and segmented by Dice-XMBD. **(D)** Scatter plots of cell fraction obtained from ground truth (*x*-axis) and five segmentation methods (*y*-axis), colored by different cell types identified from BRCA1 dataset. **(E)** Cell annotation accuracy from Dice-XMBD and other benchmarked methods in four datasets. Pairwise comparisons of Dice-XMBD and other methods: $^{*}P < 0.05$; $^{****}P < 0.0001$; n.s., not significant (Student's *t*-test).

the usability of these methods as most of them lack the ability to adapt to different channel configurations after models are trained. In this study, we develop Dice-XMBD, a generic solution for IMC image segmentation based on U-Net. Dice-XMBD overcomes the limitation of training data scarcity and achieves human-level accuracy by distilling expert knowledge from Ilastik with manual input of human as a teacher model. Moreover, by consolidating multiple channels of different proteins into two cellular structure-aware channels, Dice-XMBD provides an effective off-the-shelf solution for cell segmentation tasks across different studies without retraining that can lead to significant delay in analysis. Importantly, our evaluation results further

demonstrate Dice-XMBD's good generalization ability to predict single cells for different IMC image datasets with minimum impact to downstream analysis, suggesting its values as an generic tool for hassle-free large-scale IMC data analysis. Finally, to facilitate the analysis of large amount of IMC data currently being generated around the world, we made Dice-XMBD publicly available as an open-source software on GitHub (https://github.com/xmuyulab/Dice-XMBD).

DATA AVAILABILITY STATEMENT

All datasets used for this study can be found at GitHub (https://github.com/xmuyulab/Dice-XMBD). These datasets are downloaded from: BRCA1 (https://idr.openmicroscopy.org/search/?query=Name:idr0076ali-metabric/experimentA), BRCA2 (https://zenodo.org/record/3518284#.YLnmlS8RquU), T1D1 (https://data.mendeley.com/datasets/cydmwsfztj/1), T1D2 (part1: https://data.mendeley.com/datasets/9b262xmtm9/1, part2: https://data.mendeley.com/datasets/xbxnfg2zfs/1), respectively.

AUTHOR CONTRIBUTIONS

WY, LW, RY, and JH discussed the ideas and supervised the study. YQ and YJ implemented and conducted experiments in deep network cell segmentation. XX performed the model evaluation and biological analysis on segmentation results. XX, WY, and RY wrote the manuscript. All authors discussed and commented on the manuscript.

FUNDING

This study was funded by National Natural Science Foundation of China (grant no. 81788101 to JH).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.721229/full#supplementary-material

REFERENCES

Ali, H. R., Jackson, H. W., Zanotelli, V. R. T., Danenberg, E., Fischer, J. R., Bardwell, H., et al. (2020). Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* 1, 163–175. doi: 10.1038/s43018-020-0026-6

Andrade, A. R., Vogado, L. H., de, M. S., Veras, R., Silva, R. R., Araujo, F. H., et al. (2019). Recent computational methods for white blood cell nuclei segmentation: a comparative study. *Comput. Methods Programs Biomed.* 173, 1–14. doi: 10.1016/j.cmpb.2019.03.001

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., et al. (2014). Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* 20, 436–442. doi: 10.1038/nm.3488

Aoki, T., Chong, L. C., Takata, K., Milne, K., Hav, M., Colombo, A., et al. (2020). Single-cell transcriptome analysis reveals disease-defining t-cell subsets in the tumor microenvironment of classic hodgkin lymphoma. *Cancer Discov.* 10, 406–421. doi: 10.1158/2159-8290.CD-19-0680

Böttcher, C., van der Poel, M., Fernández-Zapata, C., Schlickeiser, S., Leman, J. K., Hsiao, C.-C., et al. (2020). Single-cell mass cytometry reveals complex myeloid cell composition in active lesions of progressive multiple sclerosis. *Acta Neuropathol. Commun.* 8, 1–18. doi: 10.1186/s40478-020-01010-8

Bouzekri, A., Esch, A., and Ornatsky, O. (2019). Multidimensional profiling of drug-treated cells by imaging mass cytometry. *FEBS Open Bio.* 9, 1652–1669. doi: 10.1002/2211-5463.12692

Brähler, S., Zinselmeyer, B. H., Raju, S., Nitschke, M., Suleiman, H., Saunders, B. T., et al. (2018). Opposing roles of dendritic cell subsets in experimental GN. *J. Am. Soc. Nephrol.* 29, 138–154. doi: 10.1681/ASN.2017030270

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., et al. (2006). Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* 7:R100. doi: 10.1186/gb-2006-7-10-r100

Catena, R., Oezcan, A., Kuett, L., Pluess, A., Schraml, P., Moch, H., et al. (2020). Highly multiplexed molecular and cellular mapping of breast cancer tissue in three dimensions using mass tomography. *bioRxiv*. doi: 10.1101/2020.05.24.113571

Chang, Q., Ornatsky, O. I., Siddiqui, I., Loboda, A., Baranov, V. I., and Hedley, D. W. (2017). Imaging mass cytometry. *Cytometry A* 91, 160–169. doi: 10.1002/cyto.a.23053

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 424–432.

Damond, N., Engler, S., Zanotelli, V. R., Schapiro, D., Wasserfall, C. H., Kusmartseva, I., et al. (2019). A map of human type 1 diabetes progression by imaging mass cytometry. *Cell Metab.* 29, 755–768. doi: 10.1016/j.cmet.2018.11.014

de Vries, N. L., Mahfouz, A., Koning, F., and de Miranda, N. F. (2020). Unraveling the complexity of the cancer microenvironment with multidimensional genomic and cytometric technologies. *Front. Oncol.* 10:1254. doi: 10.3389/fonc.2020.01254

Dey, P., Li, J., Zhang, J., Chaurasiya, S., Strom, A., Wang, H., et al. (2020). Oncogenic kras-driven metabolic reprogramming in pancreatic cancer cells utilizes cytokines from the tumor microenvironment. *Cancer Discov.* 10, 608–625. doi: 10.1158/2159-8290.CD-19-0297

Flint, L. E., Hamm, G., Ready, J. D., Ling, S., Duckett, C. J., Cross, N. A., et al. (2020). Characterization of an aggregated three-dimensional cell culture model by multimodal mass spectrometry imaging. *Anal. Chem.* 92, 12538–12547. doi: 10.1021/acs.analchem.0c02389

Giesen, C., Wang, H. A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* 11, 417–422. doi: 10.1038/nmeth.2869

Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Fullaway, C. C., et al. (2021). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *bioRxiv*. doi: 10.1101/2021.03.01.431313

Hinton, G., Vinyals, O., and Dean, J. (2015). "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop* Montreal.

Hollandi, R., Szkalisity, A., Toth, T., Tasnadi, E., Molnar, C., Mathe, B., et al. (2020). nucleAIzer: a parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Syst.* 10, 453.e6–458.e6. doi: 10.1016/j.cels.2020.04.003

Jackson, H. W., Fischer, J. R., Zanotelli, V. R., Ali, H. R., Mechera, R., Soysal, S. D., et al. (2020). The single-cell pathology landscape of breast cancer. *Nature* 578, 615–620. doi: 10.1038/s41586-019-1876-x

Jones, T. R., Carpenter, A., and Golland, P. (2005). "Voronoi-based segmentation of cells on image manifolds," in *International Workshop on Computer Vision for Biomedical Image Applications* (Berlin; Heidelberg: Springer), 535–543.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* 21, 1–35. doi: 10.1186/s13059-020-1926-6

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., et al. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 162, 184–197. doi: 10.1016/j.cell.2015.05.047

Liu, X., Song, W., Wong, B. Y., Zhang, T., Yu, S., and Lin, G. N. (2019). A comparison framework and guideline of clustering methods for mass cytometry data. *Genome Biol.* 20, 1–18. doi: 10.1186/s13059-019-1917-7

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA: IEEE), 565–571.

Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18: 35. doi: 10.1038/nri.2017.76

Potter, S. S. (2018). Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* 14, 479–492. doi: 10.1038/s41581-018-0021-7

Ramaglia, V., Sheikh-Mohamed, S., Legg, K., Park, C., Rojas, O. L., Zandee, S., et al. (2019). Multiplexed imaging of immune cells in staged multiple sclerosis lesions by mass cytometry. *Elife* 8:e48051. doi: 10.7554/eLife.48051.028

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, Munich), 234–241.

Salem, D., Li, Y., Xi, P., Cuperlovic-Culf, M., Phenix, H., and Kaern, M. (2020). Yeastnet: deep learning enabled accurate segmentation of budding yeast cells in bright-field microscopy. *bioRxiv*. doi: 10.1101/2020.11.30.402917

Schulz, D., Zanotelli, V. R. T., Fischer, J. R., Schapiro, D., Engler, S., Lun, X.-K., et al. (2018). Simultaneous multiplexed imaging of mrna and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. *Cell Syst.* 6, 25–36. doi: 10.1016/j.cels.2017.12.001

Schwabenland, M., Salié, H., Tanevski, J., Killmer, S., Lago, M. S., Schlaak, A. E., et al. (2021). Deep spatial profiling of human COVID-19 brains reveals neuroinflammation with distinct microanatomical microglia-T-cell interactions. *Immunity* 54, 1594.e11–1610.e11. doi: 10.1016/j.immuni.2021.06.002

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Sommer, C., Straehle, C., Köthe, U., and Hamprecht, F. A. (2011). "Ilastik: interactive learning and segmentation toolkit," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (Chicago, IL: IEEE), 230–233.

Stubbington, M. J., Rozenblatt-Rosen, O., Regev, A., and Teichmann, S. A. (2017). Single-cell transcriptomics to explore the immune system in health and disease. *Science* 358, 58–63. doi: 10.1126/science.aan6828

Tan, W. C. C., Nerurkar, S. N., Cai, H. Y., Ng, H. H. M., Wu, D., Wee, Y. T. F., et al. (2020). Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Commun.* 40, 135–153. doi: 10.1002/cac2.12023

Van Valen, D. A., Kudo, T., Lane, K. M., Macklin, D. N., Quach, N. T., DeFelice, M. M., et al. (2016). Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS Comput. Biol.* 12:e1005177. doi: 10.1371/journal.pcbi.1005177

Vicar, T., Balvan, J., Jaros, J., Jug, F., Kolar, R., Masarik, M., et al. (2019). Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison. *BMC Bioinformatics* 20:360. doi: 10.1186/s12859-019-2880-8

Vincent, L., and Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Comput. Arch. Lett.* 13, 583–598. doi: 10.1109/34.87344

Wang, Y. J., Traum, D., Schug, J., Gao, L., Liu, C., Atkinson, M. A., et al. (2019). Multiplexed in situ imaging mass cytometry analysis of the human endocrine pancreas and immune system in type 1 diabetes. *Cell Metab.* 29, 769–783. doi: 10.1016/j.cmet.2019.01.003

Zhang, M., Li, X., Xu, M., and Li, Q. (2018). "RBC semantic segmentation for sickle cell disease based on deformable U-Net," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 695–702.

Zhang, Y., Gao, Y., Qiao, L., Wang, W., and Chen, D. (2020). Inflammatory response cells during acute respiratory distress syndrome in patients with coronavirus disease 2019 (COVID-19). *Ann. Intern. Med.* 173, 402–404. doi: 10.7326/L20-0227

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). "Unet++: a nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Cham: Springer), 3–11.

Zrazhevskiy, P., and Gao, X. (2013). Quantum dot imaging platform for single-cell molecular profiling. *Nat. Commun.* 4, 1–12. doi: 10.1038/ncomms2635

Conflict of Interest: RY and WY are shareholders of Aginome Scientific.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 November 2021
doi: 10.3389/fgene.2021.763263

Hybrid Clustering of Single-Cell Gene Expression and Spatial Information *via* Integrated NMF and K-Means

Sooyoun Oh, Haesun Park and Xiuwei Zhang**

School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, United States

Advances in single cell transcriptomics have allowed us to study the identity of single cells. This has led to the discovery of new cell types and high resolution tissue maps of them. Technologies that measure multiple modalities of such data add more detail, but they also complicate data integration. We offer an integrated analysis of the spatial location and gene expression profiles of cells to determine their identity. We propose scHybridNMF (single-cell Hybrid Nonnegative Matrix Factorization), which performs cell type identification by combining sparse nonnegative matrix factorization (sparse NMF) with k-means clustering to cluster high-dimensional gene expression and low-dimensional location data. We show that, under multiple scenarios, including the cases where there is a small number of genes profiled and the location data is noisy, scHybridNMF outperforms sparse NMF, k-means, and an existing method that uses a hidden Markov random field to encode cell location and gene expression data for cell type identification.

Keywords: single cell transcriptomics, spatial locations, cell identity, non-negative matrix factorization, data integration

OPEN ACCESS

Edited by:
Le Ou-Yang,
Shenzhen University, China

Reviewed by:
Xingpeng Jiang,
Central China Normal University, China
Wenwen Min,
The Chinese University of Hong Kong, China

****Correspondence:***
Xiuwei Zhang
xiuwei.zhang@gatech.edu
Haesun Park
hpark@cc.gatech.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *23 August 2021*
Accepted: *13 October 2021*
Published: *08 November 2021*

Citation:
Oh S, Park H and Zhang X (2021) Hybrid Clustering of Single-Cell Gene Expression and Spatial Information via Integrated NMF and K-Means. Front. Genet. 12:763263. doi: 10.3389/fgene.2021.763263

1 INTRODUCTION

Advances in single cell RNA-Sequencing (scRNA-Seq) technology provided an unprecedented opportunity for researchers to study the identity and mechanisms of single cells (Morris, 2019). While scRNA-Seq data is a major type of data used to study single cells, it cannot fully determine the identity of a cell (McKinley et al., 2020). As such, it is important to consider other modalities such as chromatin accessibility (Cusanovich et al., 2015), protein abundance (Peterson et al., 2017), or spatial locations (Ståhl et al., 2016; Wang et al., 2018) of single cells. In particular, spatial location data can provide important information on the cells' micro-environment and cell-cell interactions (Mayr et al., 2019). In certain tissues like the brain, cells at nearby locations tend to have the same type—daughter cells tend to keep the same type and location as their mother.

Technologies that jointly profile the location and gene expression of cells are often forced to measure a small set of genes (Zhu et al., 2018). Since clustering cells using smaller gene expression profiles can be inaccurate, incorporating the cell location data can improve its accuracy. However, reconciling single cell gene expression and location data for cell type identification is challenging because different data types can have differing scales, distributions, and types of noise (Efremova and Teichmann, 2020).

Computational methods that integrate multimodal data are crucial for learning a comprehensive picture of inter- and intra-cell processes (Efremova and Teichmann, 2020; Stuart and Satija, 2019). Promising nonnegative matrix factorization (NMF) models have been

developed for cell type identification for multiple types or modalities of data (Shao and Höfer, 2017; Duren et al., 2018; Kotliar et al., 2019; Welch et al., 2019; Jin et al., 2020). However, none of these methods incorporate cell locations. On the other hand, Zhu et al. (2018) developed a HMRF (Hidden Markov Random Field) model and showed that the spatial location of cells can contribute to cell type identification.

We introduce a matrix low-rank approximation scheme, scHybridNMF (single-cell Hybrid NMF), to perform cell clustering by jointly processing cell location and gene expression data. We use a matrix low-rank approximation scheme because of the ease of preserving data characteristics through constraints and optimization terms. We combine sparse NMF with k-means clustering to cluster high-dimensional gene expression data and low-dimensional location data in an integrative way. We compare the performances of scHybridNMF against sparse NMF, k-means clustering, and HMRF on simulated and two real datasets, STARmap (Wang et al., 2018) and seqFISH+ (Eng et al., 2019), which both profile the mouse brain cortex.

2 MATERIALS AND METHODS

Matrix low-rank approximations approximate matrices as products of lower-rank matrices. Many biological clustering frameworks are designed as matrix low-rank approximation schemes because they can easily incorporate prior biological knowledge and data constraints. We formulated scHybridNMF as a combination of multiple low-rank approximations. This formulation guided the gene expression-based cell clustering with cell location information. We chose sparse NMF and k-means clustering because they could be formulated as matrix low-rank approximations, and incorporating these methods was intuitive.

2.1 Review of Sparse Nonnegative Matrix Factorization and K-Means Clustering

K-means clustering is an unsupervised learning algorithm that clusters data points by comparing pairwise distances. This metric naturally pairs with location-based data because it determines the similarity between points by how physically close they are. **Eq. 1** shows the matrix formulation for a Euclidean distance-based k-means objective for clustering $L \in \mathbb{R}^{2\times n}$, which represents location data.

$$\min_{\substack{H_L\in\{0,1\}^{k\times n}\\ H_L^T\mathbf{1}_k=\mathbf{1}_n}} \|L - W_L H_L\|_F^2, \tag{1}$$

where $\mathbf{1}_k$ and $\mathbf{1}_n$ are k- and n-length vectors of ones. The columns of $W_L \in \mathbb{R}^{2\times k}$ contain k cluster centroids, and the columns of $H_L \in \mathbb{R}^{k\times n}$ contain each point's cluster membership. If a point i belongs to a cluster j, $H_L(j, i) = 1$ and $H_L(l, i) = 0$ for $l \neq j$. The constraints preserve the hard-clustering requirement of k-means, as each data point can only belong to one cluster. This is equivalent to having one 1 per column of H_L. Additionally, k-means does not require any pre-processing, such as building a location-based neighborhood graph, on location data. Pre-processing location data may remove many of their underlying characteristics.

NMF is a dimension reduction algorithm that is well-suited for high-dimensional data. Given a nonnegative input matrix $A \in \mathbb{R}_+^{m\times n}$, NMF computes two nonnegative factors, H_A and W_A of a specified reduced dimension size k, where k is generally much smaller than m and n. The columns of $W_A \in \mathbb{R}_+^{m\times k}$ contain k cluster representatives, and the columns of $H_A \in \mathbb{R}_+^{k\times n}$ contain cluster membership information.

Sparse NMF constrains the sparsity in each column of H_A (Kim and Park, 2007). It converts the soft clustering of NMF into more of a hard clustering—a data point will have fewer nonzero entries in the cluster membership matrix and be represented by fewer cluster representatives. Sparse NMF may be interpreted as a hard clustering method if we assign each data point to the cluster of the maximal element in its column of H_A. For example, if the largest element in the first column of H_A is in the second entry, we can interpret the first data point as belonging to the second cluster.

Eq. 2 contains the formulation for sparse NMF. The first term is the objective term for standard NMF, which minimizes the difference between A and $W_A H_A$. The low-rank factors from NMF are not inherently unique, so we normalize the columns of the computed W_A and scale the rows of H_A accordingly. The second term limits the size of the elements in W_A, and the final term promotes the sparsity in each column of H_A.

$$\min_{\{W_A,H_A\}\geq 0} \|A - W_A H_A\|_F^2 + \beta\|W_A\|_F^2 + \gamma\sum_{i=1}^{n}\|H_A(:,i)\|_1^2. \tag{2}$$

2.2 Multimodal Objective

Let $A \in \mathbb{R}_+^{m\times n}$ denote the normalized gene expression matrix and $L \in \mathbb{R}^{2\times n}$ denote the two-dimensional cell location coordinates, where m is the number of genes and n is the number of cells. To get the normalized gene expression matrix, we first scaled the rows of the raw count matrix, $\tilde{A}$, by its library size, then set $A = \log_2(\tilde{A} + 1)$. We computed W_A and H_A from sparse NMF on the gene expression data and W_L and H_L from k-means clustering on the location data. We used the same k in both methods, which allowed for a direct comparison between the two data types. We assumed that k is already known for each dataset. **Eq. 3** is the objective function for the multimodal clustering:

$$\min_{\{W_A,H_A\}\geq 0} g(W_A, H_A) = \min_{\{W_A,H_A\}\geq 0} \|A - W_A H_A\|_F^2 + \alpha\left\|H_A - H_A \circ \hat{H}_L\right\|_F^2. \tag{3}$$

In **Eq. 3**, $\circ$ represents the element-wise product between two matrices, and the second term forms the consensus between the clustering results from sparse NMF and k-means clustering. $\hat{H}_L$ was obtained by converting H_L into a matrix of confidence scores that considered how close each cell was to the edge of its location-based cluster. We found the index of two closest cluster centroids to each cell i, then assigned values to entries in $\hat{H}_L$ (**Eq. 4**). All

other entries of $\hat{H}_L$ remained zero. As such, we compared H_A with $\hat{H}_L$, and not with H_L directly.

$$\hat{H}_L(j,i) = \begin{cases} \frac{\|W_L(:,j) - L(:,i)\|_2}{\sum_{j'=1}^{2} \|W_L(:,j') - L(:,i)\|_2}, & \text{if } j \text{ is one of the top 2 cluster indices for cell } i. \\ 0, & j \text{ for all other clusters.} \end{cases} \tag{4}$$

Instead of forcing H_A and $\hat{H}_L$ to be similar overall, the second term in **Eq. 3** forced H_A and $\hat{H}_L$ to be similar in terms of their cluster memberships. In other words, the second term of **Eq. 3** aimed to match the location of the largest element in each column of H_A and the location of the two nonzero elements in the corresponding column of $\hat{H}_L$.

The main focus of this work was to use cell location information to aid the gene expression-based clustering of cells. Because we specifically adapted gene clusters to incorporate location cluster information, our design sought to align the cluster membership matrices while still considering the accuracy of the gene expression clustering. We did not include a sparsity term for H_A, the final cluster membership matrix, because imposing the sparsity terms may eliminate nuance in the integration of both clustering schemes, and thus result in a loss of information that could better serve to cluster the cells.

2.3 Proposed Algorithm

scHybridNMF optimized **Eq. 3** to combine the clusters of sparse NMF on A and k-means on L. To get the initial H_A for the consolidated algorithm, we ran sparse NMF on A. We then computed k-means clustering on L. We computed initial centroids by taking the means of each cell's locations within the gene expression-based clusters.

scHybridNMF used block coordinate descent for computing H_A and W_A. These two terms were computed via an alternating nonnegative least squares (ANLS) formulation.

$$\left\|H_A - H_A \circ \hat{H}_L\right\|_F^2 = \left\|H_A \circ \mathbf{1}_{k\times n} - H_A \circ \hat{H}_L\right\|_F^2 = \|H_A \circ C\|_F^2, \tag{5}$$

where $C = \mathbf{1}_{k\times n} - \hat{H}_L$ and $\mathbf{1}_{k\times n}$ is a $k \times n$ matrix of ones. We represented the element-wise product in a block-ANLS formulation by computing it column-by-column. Column i of H_A is updated as follows:

$$H_A(:,i) \leftarrow \underset{H_A(:,i)\geq 0}{\arg\min} \left\| \begin{pmatrix} W_A \\ \sqrt{\alpha} * \mathrm{diag}(C(:,i)) \end{pmatrix} H_A(:,i) - \begin{pmatrix} A(:,i) \\ \mathbf{0}_k \end{pmatrix} \right\|_F^2, \tag{6}$$

where $i \in \{1, \ldots, k\}$, $\mathbf{1}_k$ is a k-length vector of ones, and $\mathbf{0}_k$ is a k-length vector of zeros. Each column in H_A was element-wise multiplied to each column in C in **Eq. 5**, which can be represented as a left-multiplication of the column of H_A by a matrix whose diagonal entries are the corresponding column of C.

For W_A, we used the following update rule:

$$W_A \leftarrow \underset{W_A\geq 0}{\arg\min} \|A - W_A H_A\|_F^2. \tag{7}$$

The overall scheme is described in **Algorithm 1**. There exist many stopping criteria that can be used. We used two: a maximum number of iterations and a normalized KKT condition residual check, as used in SymNMF (Kuang et al., 2015).

Algorithm 1. scHybridNMF: an algorithm to minimize **Eq. 3**

Input : normalized gene expression matrix $A \in \mathbb{R}_+^{m\times n}$, cell location matrix $L \in \mathbb{R}^{2\times n}$, number of clusters k.
Compute H_A using Eqn. (2);
Compute initial location centroids using cluster labels from H_A;
Compute W_L, H_L using Eqn. (1) and initial computed centroids;
Compute $\hat{H}_L$ using Eqn. (4);
$C = \mathbf{1}_{k\times n} - \hat{H}_L$;
while *stopping criterion has not been met* **do**
  Compute W_A using Eqn. (7);
  for $i = 1, \ldots, n$ **do**
    Compute $H_A(:,i)$ using Eqn. (6);
Output : W_A and H_A.

2.4 Parameters

In line 1 of **Algorithm 1**, we computed sparse NMF on the data matrix A through **Eq. 2**. This formulation involved β and γ, which controlled the size of the entries of W_A and the sparsity of H_A, respectively. To ensure that the last two terms were proportionate to the first term in the formulation, we formulated β and γ to have a denominator of $\|A\|_F^2$, which is the maximum value the first term can take. We also formulated the parameters based on the dimensions of W_A and H_A. We set the numerator of β to be m, which is the number of rows of W_A, and we set the numerator of γ to be n, which is the number of columns of H_A. The final formulations were $\beta = \frac{m}{\|A\|_F^2}$ and $\gamma = \frac{n}{\|A\|_F^2}$.

The parameter α in the hybrid clustering scheme was designed to control the degree to which the consensus clustering was influenced by the location-based clusters. The maximum number of iterations to run the main BCD was set to be 500 so it is not triggered as much as the other stopping criterion. The tolerance level, *tol*, of the normalized KKT residual check had a default value of 0.01. The relationship between α and *tol* is interesting. A smaller α, which prioritizes gene expression-based clusters, required a larger *tol*, as scHybridNMF's clusters did not converge otherwise. Likewise, a larger α, which prioritizes cell location-based clusters, required a smaller *tol* to ensure that scHybridNMF did not return the same clusters as k-means. For α and *tol*, we recommend using values between 0 and 1.

2.5 Convergence of Algorithm

We used a block coordinate descent (BCD) framework to optimize **Eq. 3**. BCD solves subgroups of problems for a set of variables of interest, which iteratively minimizes the total objective function. We used the minimization version of the two-block BCD method, which assigned $H_A^{(j)}$ and $W_A^{(j)}$ values that minimized **Eq. 3** one-at-a-time.

An important theorem regarding BCD states that if a continuously differentiable function over a set of closed convex sets is minimized by BCD, every limit point obtained from uniquely minimizing the subproblems in BCD is a stationary point (Bertsekas et al., 1997). This theorem has the additional property that the uniqueness of the minimum is not necessary for a two-block BCD nonlinear minimization scheme (Grippo and Sciandrone, 2000). This was used to show the convergence of a two-block formulation for solving regular NMF *via* ANLS (Kim et al., 2014).

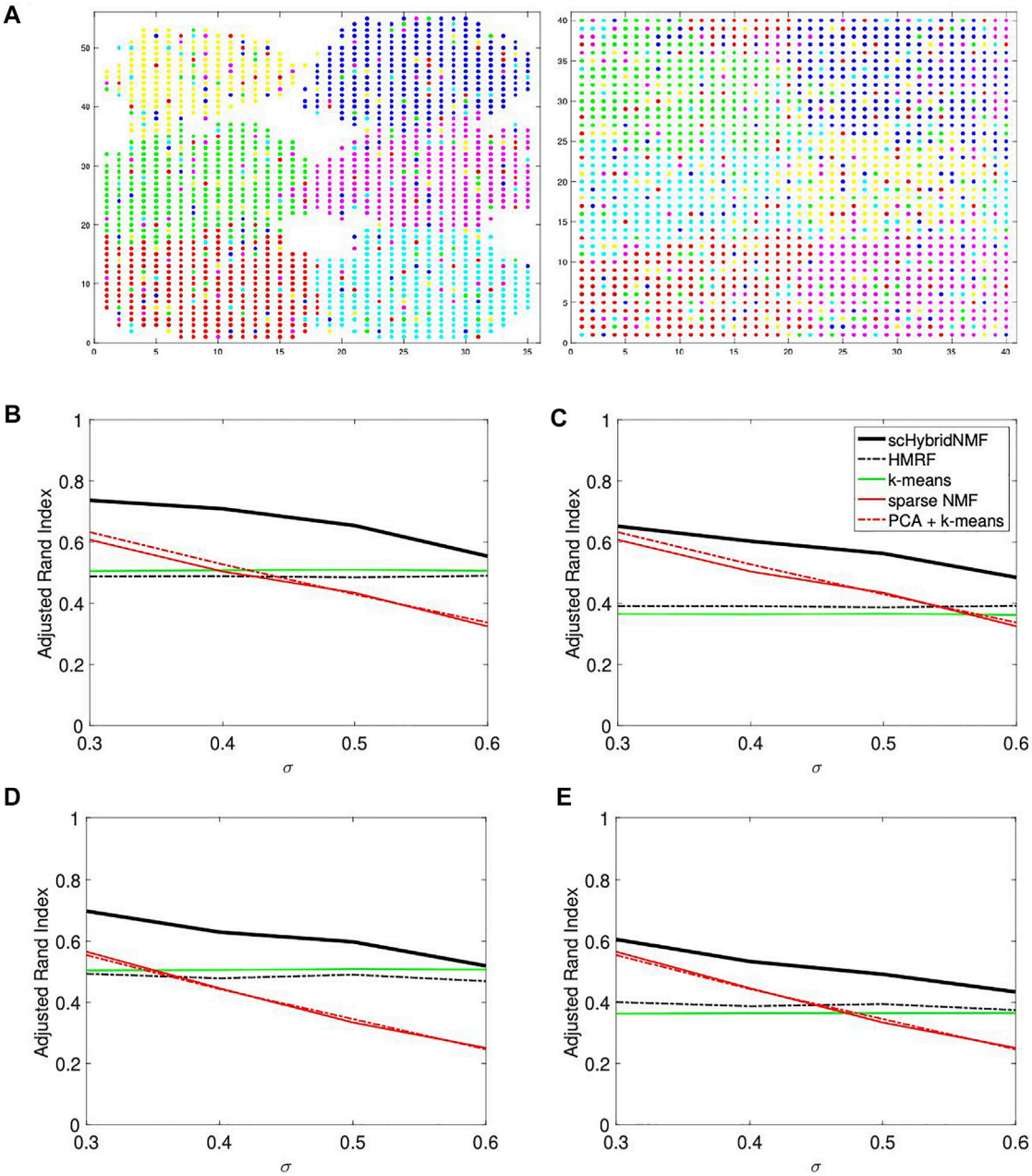


FIGURE 1 | (A) An example of noise in location data. The data had $\sigma = 0.3$ and 20% noise. In each plot, there are six point colors that correspond to true cluster labels. Left: strong spatial patterns; right: weak spatial patterns. Note that certain cell types are not contiguous in the right plot. **(B–E)** Performance vs sigma plots for location data with strong spatial patterns. Each plot shares the same legend as plot **(C)**. **(B)** No sampling, 20% noise. **(C)** No sampling, 30% noise. **(D)** 50% sampling, 20% noise. **(E)** 50% sampling, 30% noise.

Given the constrained nonlinear minimization objective in **Eq. 3**, we rewrote the block coordinate descent as two ANLS formulations, which follow from **Eq. 6** and **Eq. 7**:

$$H_A(:,i)^{(j)} \leftarrow \underset{H_A(:,i)\geq 0}{\arg\min} \left\| \begin{pmatrix} W_A^{(j-1)} \\ \sqrt{\alpha} * \mathrm{diag}(C(:,i)) \end{pmatrix} H_A(:,i) - \begin{pmatrix} A(:,i) \\ \mathbf{0}_k \end{pmatrix} \right\|_F^2, \tag{8}$$

$$W_A^{(j)} \leftarrow \underset{W_A\geq 0}{\arg\min} \left\| \left(H_A^{(j)}\right)^T W_A^T - A^T \right\|_F^2. \tag{9}$$

Eqs. 8 and **9** were executed iteratively to solve for H_A and W_A. We considered **Eq. 8** to be one block calculation for the entire H_A matrix because the calculation of a column of $H_A^{(j)}$ does not involve any other column. **Eqs. 8** and **9** constituted a valid minimization scheme equivalent to minimizing **Eq. 3**. As such,

the theorem by Bertsekas is applicable to this two-block BCD scheme for solving scHybridNMF (Bertsekas et al., 1997; Kim et al., 2014):

THEOREM 1 Every limit point $\{W_A^{(j)}, H_A^{(j)}\}$ *calculated iteratively* via ***Eqs. 8–9*** *is a stationary point of* ***Eq. 3***.

3 RESULTS

We tested the performance of scHybridNMF against simulated and real data. For real data, we experimented on the STARmap and seqFISH+ datasets, both of which catalogue the mouse brain cortex (Eng et al., 2019). For STARmap, we compared against sparse NMF and k-means clustering to show an improvement of our hybrid scheme over each method. For the simulated data and seqFISH+, we also compared against HMRF (Zhu et al., 2018), a method that also performs consensus cell clustering on gene expression and cell location data. HMRF models cell locations as nodes on a graph, where cells are connected if they are neighbors in location. It clusters cells by searching for coherent gene expression patterns within neighboring cells.

We implemented the code in MATLAB 2019b. For sparse NMF, we used MATLAB code presented by Kim and Park (Kim and Park, 2008). All experiments were executed on a computer with 2.4 GHz 8-Core Intel Core i9 and 32 GB 2400 MHz DDR4 RAM.

3.1 Simulated Data

We used SymSim to simulate single cell gene expression data, where each cell has one of six cell types (Zhang et al., 2019). Each dataset has 1,600 cells and 600 genes. We developed two types of cell location datasets, where one has strong and the other has weak spatial patterns. For each case, we generated location data with 20 and 30% noise by randomly choosing 20 and 30% of the cells and assigning them to locations outside of their original cell type cluster. Adding noise to the locations made the data more realistic. **Figure 1A** shows an example of location data with 20% noise.

SymSim has a parameter σ that adjusts the within-cluster heterogeneity of gene expression. When σ increased, the gene expression-based clusters were less separable, and gene expression-based clustering algorithms were less reliable. We used σ = (0.3, 0.4, 0.5, 0.6). For each sigma, 10 gene expression-cell location datasets were generated. For each location matrix, we generated 10 noisy location datasets per noise level.

Many current technologies, especially image-based technologies that pairwise measure the gene expression and spatial locations of single cells, cannot also sequence many genes (Zhu et al., 2018; McKinley et al., 2020). To mimic the limitations of current technology, we additionally created gene-sampled data by randomly sampling 50%, or 300, of the genes from each of the original gene expression datasets.

We compared the quality of clusters determined by gene expression clustering, cell location clustering, and hybrid clustering. The methods we used for gene expression clustering were sparse NMF and PCA plus k-means clustering, which provided a baseline for the performance of sparse NMF. For example, a poor performance from PCA plus k-means clustering justified similarly poor performance of sparse NMF. For location-based clustering, we used k-means clustering. To cluster both data types, we used scHybridNMF and HMRF. HMRF uses a parameter, called beta, which accounts for smoothness. We determined the performance of HMRF as the average performance across 5 values, (0, 20, 40, 60, 80), for beta.

We calculated the adjusted Rand index (ARI) between the calculated and ground truth clusters for each clustering method across each experiment. ARI quantifies how similar two clustering schemes are. If a clustering is very similar to the ground truth clustering, the ARI should be close to 1. We used the sparse NMF and k-means clustering that were used in the steps of **Algorithm 1** to calculate their respective ARI values.

3.1.1 Location Data With Strong Spatial Patterns

The location data with strong spatial patterns had significant spatial gaps between clusters (**Figure 1A**, left plot), and k-means clustering did well separating clusters. For these cases, location clustering played a major role in the multimodal clustering scheme. For σ = (0.3, 0.4, 0.5, 0.6), we used α = (50, 55, 60, 60) and *tol* = (0.02, 0.02, 0.02, 0.04). We used the same parameters for data with and without gene sampling. We plotted the average ARIs as a function of σ in **Figures 1B–E**. **Figures 1B,C** show the ARIs for data with no gene sampling, and **Figures 1D,E** show the ARIs for data with 50% gene sampling.

The plots showed a clear improvement of scHybridNMF over every other method. scHybridNMF followed the same performance trend as gene expression-based clustering across each σ. In contrast, HMRF's performance over every σ value was constant. This was highly similar to the performance of location-based clustering, which was often outperformed by gene expression clustering.

3.1.2 Location Data With Weak Spatial Patterns

In this location data, the boundaries between clusters were hard to determine (**Figure 1A**, right plot). As such, k-means clustering experienced more difficulty, and gene expression information was more useful in the multimodal clustering scheme. For σ = (0.3, 0.4, 0.5, 0.6), we used α = (0.015, 0.02, 0.025, 0.04) and *tol* = (0.2, 0.2, 0.2, 0.2). We used the same parameters for data with and without gene sampling. We plotted the average ARIs as a function of σ in **Figures 2A–D**. **Figure 2A,B** show the ARIs for data with no gene sampling, and **Figures 2C,D** show the ARIs for data with 50% gene sampling.

scHybridNMF and HMRF had the same performance trends as they did in **Figures 1B–E**. However, neither the gene expression nor the cell location data accurately represented the underlying data well—the ARIs and qualities of the gene expression- and location-based clusterings for larger σ were very low. Because scHybridNMF drew information from these clusters, it was difficult to gain significantly better information than what was found individually.

scHybridNMF still maintained higher levels of performance in most cases. When σ increased, the clusters were less separable with gene expression data, and the performance of sparse NMF decreased. This caused the decrease of the performance of

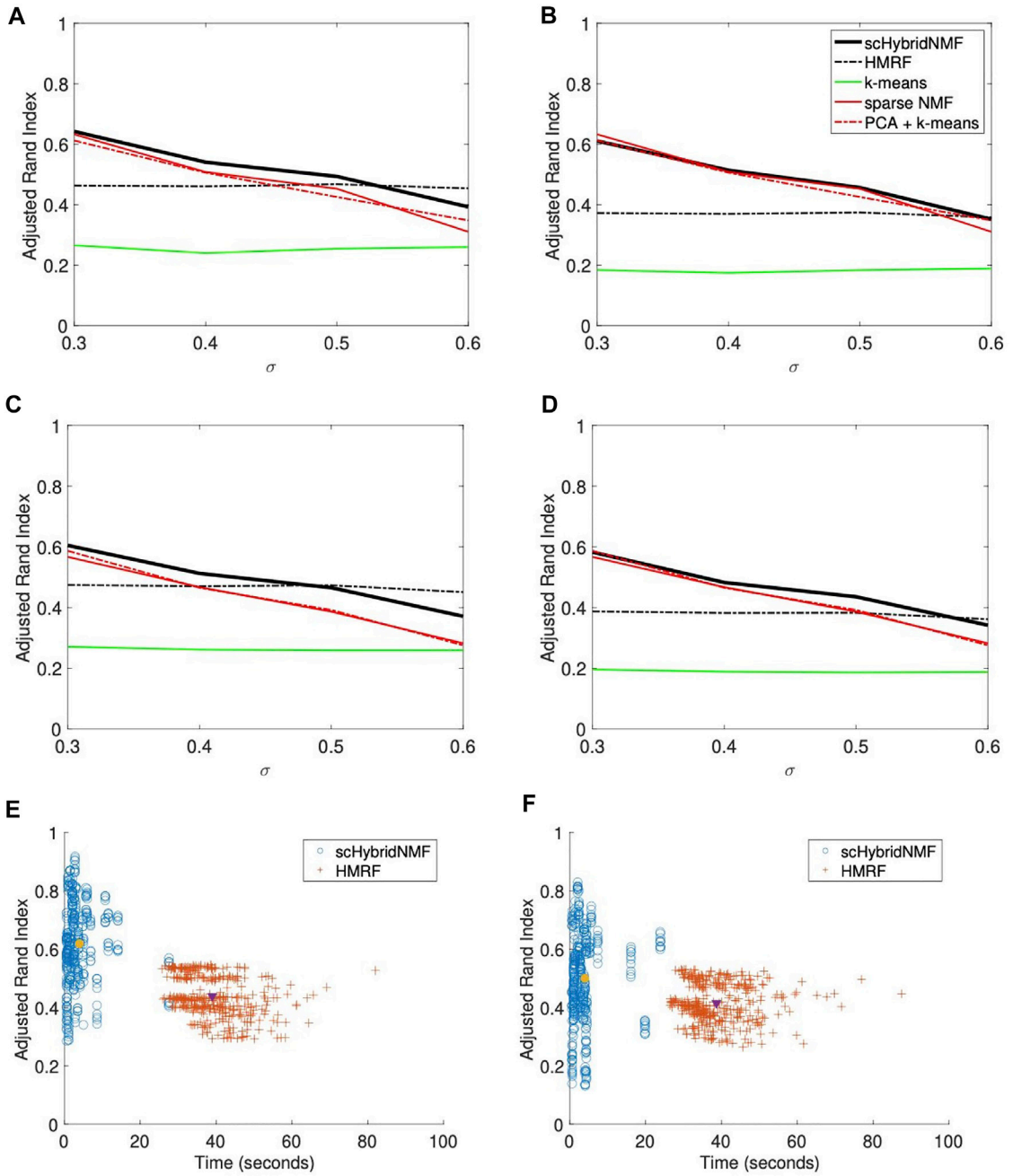


FIGURE 2 | (A–D) Performance vs sigma plots for location data with weak spatial patterns. Each plot shares the same legend as plot **(B)**. **(A)** No sampling, 20% noise. **(B)** No sampling, 30% noise. **(C)** 50% sampling, 20% noise. **(D)** 50% sampling, 30% noise. **(E,F)** Time vs performance dot plots of scHybridNMF and HMRF on gene expression data with no gene sampling and location data with strong and weak spatial patterns. **(E)** Strong spatial patterns. **(F)** Weak spatial patterns.

scHybridNMF. Although it did not perform very well with small σ, the performance of HMRF was not affected much by the increase of σ, and it started to decrease only when $\sigma > 0.5$. This was likely due to the fact that the neighborhood graph approach used in HMRF is good at learning from location data. However, as evidenced by the performance patterns of HMRF across different σ values, HMRF is not able to make full use of high-quality gene expression data.

3.1.3 Timings

We presented two separate dot plots of algorithm completion time vs ARI for each data matrix pair with no gene sampling (**Figures 2E,F**). An ideal algorithm would have most points in the top-left of the plot; these points correspond to high ARIs with smaller completion times. To show overall trends, we consolidated the noise levels for each plot. For HMRF, we timed from creating the graphical representation to the end

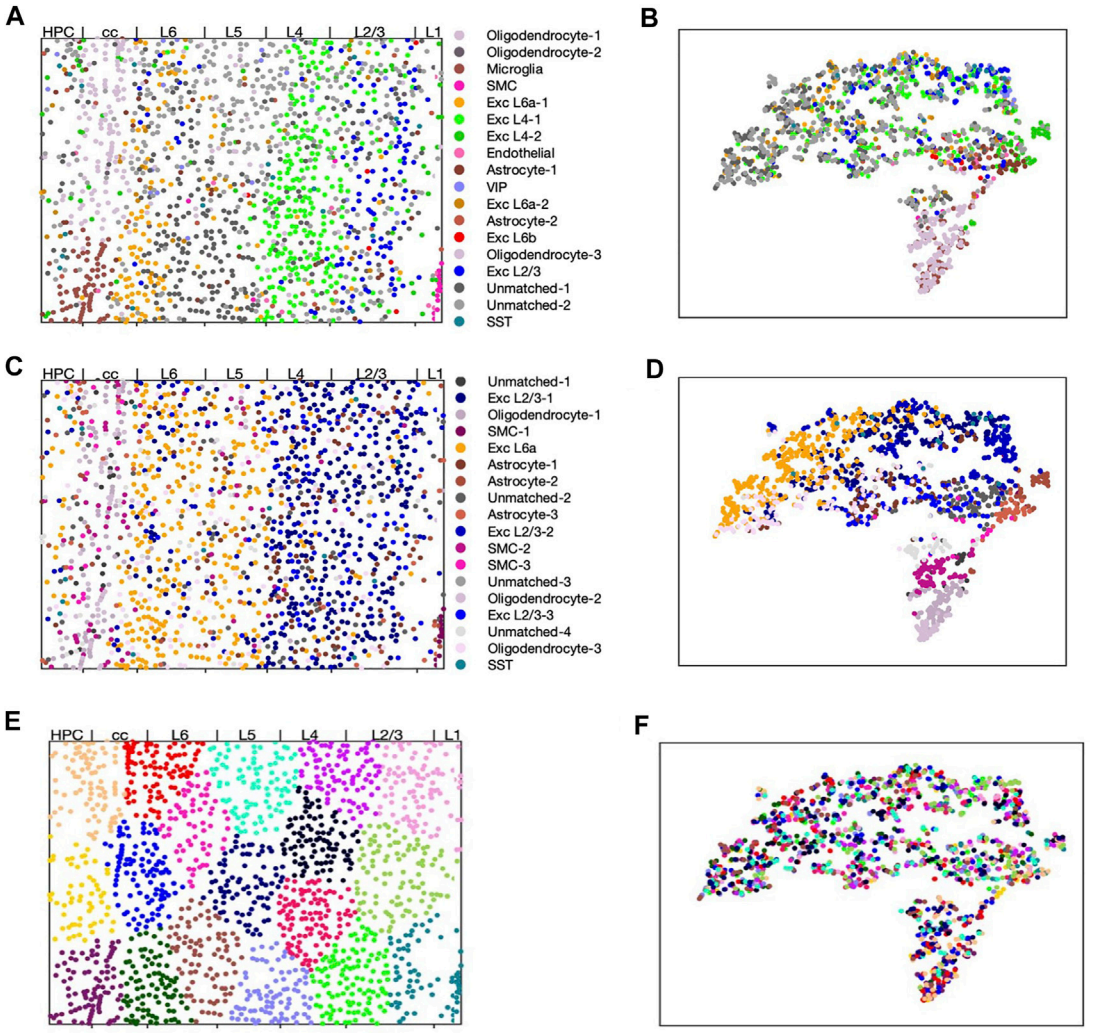


FIGURE 3 | scHybridNMF, sparse NMF and k-means clustering on STARmap data. The layers are labelled by Wang et al. (2018). **(A,C,E)** Visualizing cells in spacial location with cell cluster labels from respectively scHybridNMF, sparse NMF and k-means. **(B,D,F)** cells shown in t-SNE plots of gene expression colored with cluster labels from respectively scHybridNMF, sparse NMF and k-means.

for each parameter choice, then averaged the times. For scHybridNMF, we timed from computing sparse NMF to the end. Both algorithm timings matched the values used to compute the ARI values in **Figures 1B–E** and **Figures 2A–D**. **Figure 2E** shows the time and performance data of each point represented in **Figures 1B,C**, and **Figure 2F** shows the time and performance data of each point represented in **Figures 2A,B**.

These experiments showed that scHybridNMF performed well with varying levels of gene sampling and location noise. The fact that scHybridNMF consistently outperformed sparse NMF and k-means indicates that it is likely to be successful on real data.

3.2 STARmap Dataset

Wang *et al* developed STARmap, which profiled both "thin" and "thick" cross-sections in the mouse brain cortex (Wang et al., 2018). We used the "thin" dataset, which profiled from layer 1 of the cortex to some of the hippocampus. This dataset has 1,549 cells and 1,020 genes. The cell types noted by Wang et al. (2018) had distinct patterns in their gene expression, cell location, or a combination of both. For example, excitatory neurons may have subtypes specific to certain cortex layers (Tasic et al., 2016). These can be identified by their presence in one or two layers of the cortex, but they are harder to differentiate using only gene expression.

We compared scHybridNMF against sparse NMF and k-means clustering to show that it recovered underlying information that could not be recovered using only one modality of data. We used $k = 18$, which is the same k used by Wang et al. (2018). The final clusters we profiled for k-means and sparse NMF were the clusters used as input to scHybridNMF. For scHybridNMF, we set $\alpha = 0.015$ and $tol = 0.1$. This was because the location data was not very separable.

To better compare our clustering results against the underlying cell types, we assigned cell type labels to clusters. We used Scran, a program that detects differentially-expressed (DE) genes given clusters, to find the top 20 such genes per cluster (Lun et al., 2016). We then assigned cell type labels by measuring the overlap of DE genes and marker genes for known cell types in the STARmap data (Wang et al., 2018). The final cluster labels are shown in **Supplementary Table S1**.

We visualized the clustering results in **Figure 3**. We first split the different possible cluster colors by the different cell types found, with a particular effort given towards making the excitatory neuron subtype colors distinct. We then consolidated clusters that shared the same cluster label, then assigned them different shades of the color that defined the shared cell type label.

We found that none of the clusters found by k-means clustering matched any known cell types (**Figure 3E,F**). Using a location-based clustering method only finds clusters based on the location density pattern and the intrinsic characteristics of the clustering method. Therefore, with this STARmap dataset, k-means clustering found similarly-sized and shaped structures that separated the locations evenly. scHybridNMF, on the other hand, found clusters with the striped structures of the layers of the cortex while also recovering cell types that were less spatially conserved (**Figure 3A,B**).

We performed comprehensive comparison between the results of sparse NMF and scHybridNMF. As a preliminary measure, we computed the ARI between the clusters determined by Wang et al. (2018), noted as ground truth clusters, and the clusters from scHybridNMF and sparse NMF. (Wang et al., 2018). provided labels for 1,389 cells, and we further removed from consideration the cells that Wang et al. (2018) excluded from clustering. This left a total of 1,207 cells for ARI calculation. We found that the ARI between the ground truth and sparse NMF's clusters to be 0.255, and the ARI between the ground truth and scHybridNMF's clusters to be 0.21. Sparse NMF's marginally higher ARI and better-clustered tSNE visualization of gene expression data (**Figure 3D**) can be explained by the fact that the cluster annotations by Wang et al. (2018) were determined through just the gene expression matrix. However, the spatial distribution of the clusters determined by scHybridNMF better fit the shape of the layer-specific regions in the ground truth labels than the clusters determined by sparse NMF (**Figures 3A–D**). As such, we further examined both the spatial and gene expression components of the cell type annotations.

Most of the clusters recovered by sparse NMF were similar to those found by scHybridNMF, but scHybridNMF was able to recover major cell types that sparse NMF was not able to (**Figures 3A–D**). These cell types were separable by gene expression, but were more clearly separated by locations. scHybridNMF was able to recover distinct L2/3, L4, and L6a excitatory neurons, while sparse NMF was not.

3.2.1 scHybridNMF Separates Different Types of Excitatory Neurons

Excitatory neurons have layer-based subtypes (Tasic et al., 2016). These subtypes differ in their locations and gene expression profiles, and each have their own marker genes (Tasic et al., 2016; Wang et al., 2018). Here, we show that scHybridNMF better isolated three subtypes of excitatory neurons, L2/3, L4 and L6a, than sparse NMF.

In **Figure 4A,B**, we highlighted the clusters relevant to L2/3, L4 and L6a excitatory neurons while keeping other clusters in grey. We observed two separate clusters with scHybridNMF in the upper layers of the brain cortex that corresponded to L2/3 and L4 excitatory neurons (blue and pink clusters in **Figure 4A**, **Supplementary Table S1**). In contrast, sparse NMF was not able to detect two clear clusters for L2/3 and L4 excitatory neurons. In fact, there were no cluster found by sparse NMF that could be mapped to L4 excitatory neurons (**Supplementary Table S1**). Additionally, the clusters that were annotated as L6a excitatory neurons in each method had very different location distributions (**Figure 4A,B**). Compared to sparse NMF, the cell types annotated by the scHybridNMF clustering were more in line with the layer structure.

We then investigated whether the expression of marker genes supported the clustering by scHybridNMF. We examined Lamp5, Nrsn1, and Rprm, which are noted by (Wang et al., 2018) to be marker genes for L2/3, L4, and L6a excitatory neurons. First, we showed that the expression level of these genes exhibited the spatial pattern of the corresponding layer (**Supplementary Figure S1**). Then, we compared the differential expression of these genes across scHybridNMF and sparse NMF clusters, shown in box plots in **Figures 4C–E**.

We used normalized, log-transformed gene expressions to create box plots of the genes across each cluster. Clusters 15 and 6 of scHybridNMF, which were annotated as L2/3 and L4 excitatory neurons, distinctly exhibited higher expressions of Lamp5 and Nrsn1. This differentiation supported the location-based separation of the two excitatory neuron subtypes. On the other hand, for sparse NMF, clusters 2 and 15 had a highly differential level of expression of Lamp5 in **Figure 4C**. However, the clusters that exhibited high levels of Nrsn1 were also clusters 2 and 15, which were labeled as L2/3 excitatory neurons during the annotation procedure (**Figure 4D**). The third sparse NMF cluster annotated as L2/3 excitatory neurons, cluster 10, did not exhibit differential expression of these genes (**Figure 4C,D**).

We additionally observed that scHybridNMF was better able to recover L6a excitatory neurons than sparse NMF. L6a excitatory neurons highly expressed Rprm, were located in the deeper parts of the cortex, and were arranged in a layer-like structure (**Supplementary Figure S1**). Cluster 5 from both scHybridNMF and sparse NMF corresponded to L6a excitatory neurons (**Supplementary Table S1**). Cluster 5 of scHybridNMF showed a more distinct expression of Rprm compared to cluster 5 of sparse NMF (**Figure 4E**). Its spatial pattern, in **Figure 4A**, also more closely matched the spatial pattern of the cells that highly exhibited Rprm.

It is worth noting that the cell type annotations obtained in **Supplementary Table S1** were based on multiple marker genes per cell type. For example, we additionally found that Nrep and Zmat4, noted by (Wang et al., 2018) to be marker genes for L4 excitatory neurons, exhibited the same differential expression for cluster 6 of scHybridNMF. Overall, we showed that scHybridNMF found excitatory neuron subtypes better than sparseNMF in terms of both cell locations and marker gene expression levels.

3.3 seqFISH+ Dataset

Eng et al. (2019) profiled the mouse brain cortex and sub-ventricular zone (SVZ) across 7 fields of view (FOV) using the

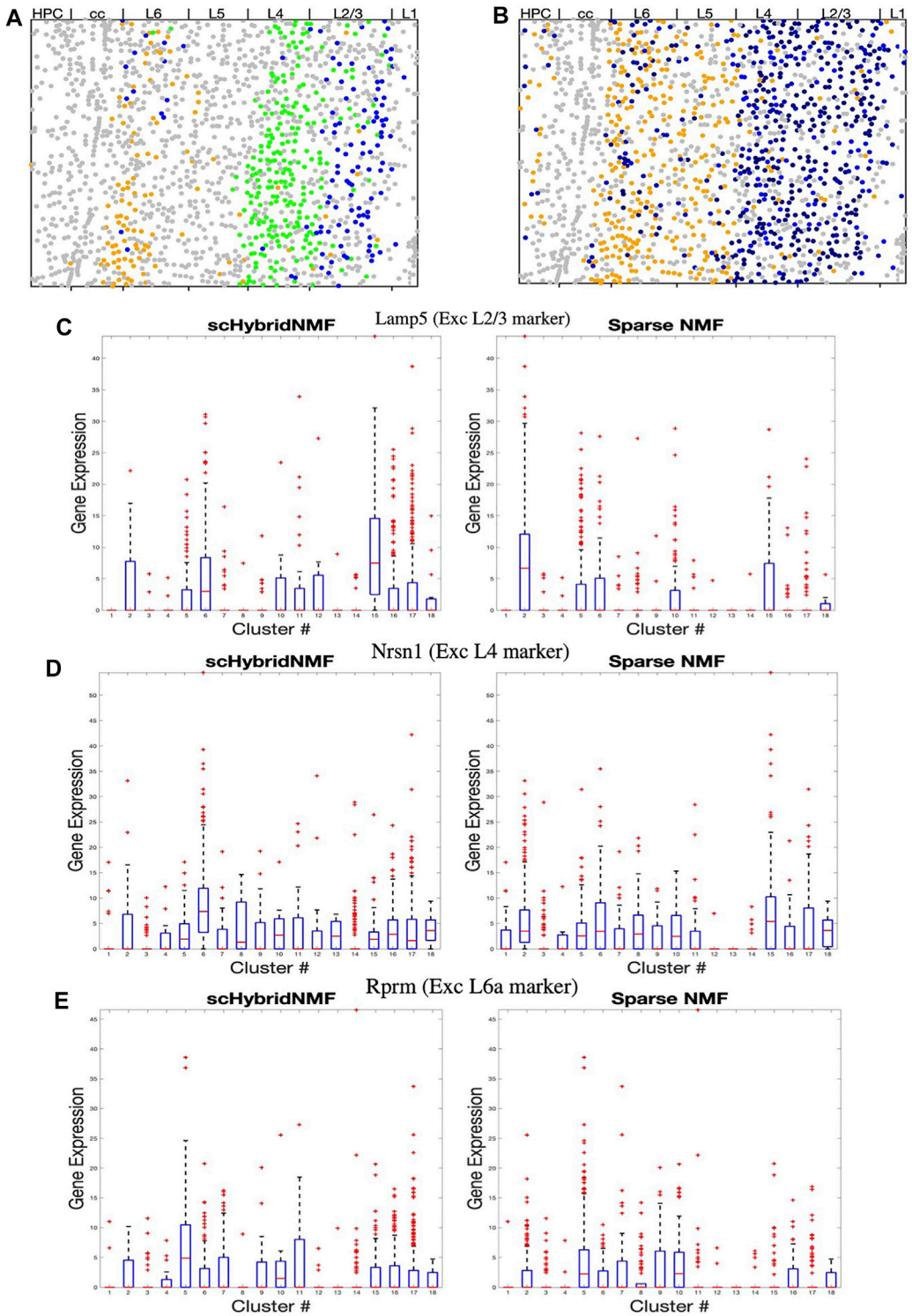


FIGURE 4 | (A–B) Dot plots of clusters that best match L2/3, L4, and L6a excitatory neurons. All other clusters are in grey. **(A)** Cluster 5 (L6a, orange), 6 (L4, green), and 15 (L2/3,blue) from scHybridNMF. **(B)** Clusters 5 (L6a, orange), 15 (L2/3, black), and 2 (L2/3, black) from sparse NMF. **(C–E)** Box plots of the expressions of Lamp5, Nrsn1, and Rprm across each cluster.

seqFISH+ technique. Five of the FOV were taken from the visual cortex, and 2 from the SVZ. We analyzed 523 cells in the 5 visual cortex FOVs, which encompassed cells from L1 to L6. The gene expression levels of 10,000 genes and locations were profiled for each cell. We computed the means and standard deviations of each gene's expression levels across each cell, and we kept the

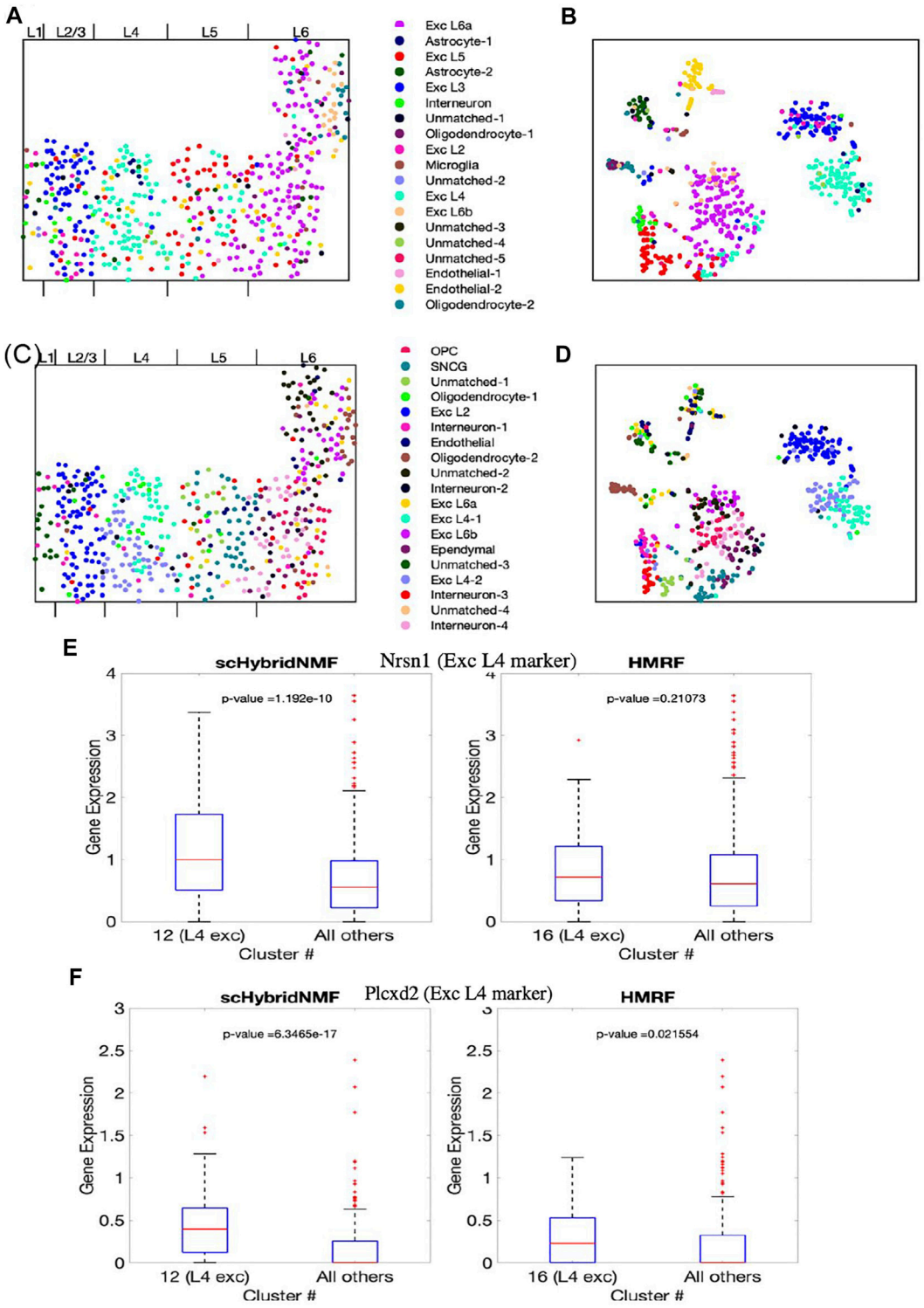


FIGURE 5 | The clustering results of scHybridNMF and HMRF on seqFISH+ data. **(A,C)** Cells visualized in spatial locations with clustering labels from respectively scHybridNMF and HMRF. **(B,D)** Cells visualized in t-SNE plots of gene expression with clustering labels from respectively scHybridNMF and HMRF. The cluster labels are shown in the middle. The layers are labelled by Dries et al. (2021). **(E,F)** Box plots of the expressions of Nrsn1 and Plcxd2 in cells in exc L4 clusters vs all other cells. *p*-values were calculated with a two-sample *t*-test that tested if the population mean of exc L4 clusters were larger than that of the rest of the cells.

genes with means greater than 0.7 and correlations of variation greater than 1.2. This left 1,047 genes. We then added all of the marker genes from Tasic et al. (2016) that were not already in the set of 1,047 genes, which resulted in a total of 1,198 genes.

We set the number of clusters, k, to be 19. The labels for the original seqFISH+ dataset were derived from the 49 transcriptomic cell types identified by Tasic et al. (2016). By grouping together cell types in the minor 49, we found 20 cell types. We then explored different numbers of clusters around 20, and found that $k = 19$ gave the most intriguing results. For scHybridNMF, we set $\alpha = 45$ and used a tolerance of 0.05. For the HMRF algorithm, we used a beta value of 10, which was the beta value that gave clusters that were the most consistent with the underlying anatomical structure of the visual cortex.

We used Scran to find the top 20 DE genes per cluster (Lun et al., 2016). We then cross-referenced these with marker genes found by Tasic et al. (2016) and Eng et al. (2019) to map the clusters to tentative cell types. However, certain cell types from Eng et al. (2019) did not match the actual cell locations within the brain cortex. For example, cells annotated as layer 2 excitatory neurons seemed to reside in deeper cortex layers. As such, we considered the location-specific cell type information provided by Tasic et al. (2016) with a higher degree of confidence, and did not compute the ARI with the labels provided by Eng et al. (2019).

The final cluster labels are shown in **Supplementary Table S2**. We visualized the cluster results of scHybridNMF and HMRF on the cell location and gene expression spaces (**Figures 5A–D**). We again split the different possible cluster colors by the different labels, with a particular effort given towards making the excitatory neuron subtype colors distinct. We then consolidated clusters that shared the same cluster label, then assigned them different shades of the color that defined the shared cell type label.

As a preliminary reference, we calculated the Silhouette values of the clusterings found by scHybridNMF and HMRF for gene expression values. However, both methods had very similar performances across every cluster found, even clusters that were left unmapped. As such, we conducted a gene ontology (GO) term analysis for the DE genes found by Scran.

3.3.1 scHybridNMF Detects L4 Excitatory Neurons

Layer-specific excitatory neurons form contiguous, column-like structures, and they also have unique gene expression profiles (Tasic et al., 2016). The Giotto authors labelled distinct physical layers, numbered 1, 2/3, 4, 5, and 6, in the seqFISH+ dataset (Dries et al., 2021). We found that there were excitatory neuron subtypes that generally corresponded to each of layers 2/3 to 6. In particular, we found that scHybridNMF was able to recover a cluster (cluster 12 in **Supplementary Table S2**) that better corresponded to L4 excitatory neurons than HMRF's cluster (cluster 16 in **Supplementary Table S2**).

To further investigate this, we looked into the expressions of marker genes, especially Nrsn1 and Plcxd2. Nrsn1 was noted by Eng et al. (2019) to be a marker gene for excitatory neurons, and is visibly highly expressed in layer 4 of the cortex. Plcxd2 is shown by (Wang et al., 2018) to be a marker gene for neuronal cells, especially L4 and L5 excitatory neurons, but we show that in the seqFISH+ dataset, this is uniquely highly expressed in layer 4. All other marker genes are shown in **Supplementary Figures S2,S3**.

First, we saw that the cells that highly expressed these genes were grouped together in a layer-like shape (**Supplementary Figures S2A,B**), confirming the marker genes' spatial patterns. We then visualized the different marker gene expressions with box plots, comparing the expressions within L4 excitatory neuron clusters of scHybridNMF and HMRF against the rest of the cells (**Figures 5E,F**). We found that, with a threshold of $p < 0.01$, cluster 12 of scHybridNMF exhibited a significantly higher expression of Nrsn1 and Plcxd2 than the rest of the cells (**Figures 5E,F**). In contrast, HMRF failed to reject the null hypothesis, with p-values of 0.21 and 0.02.

3.3.2 Layer 6b Excitatory Neurons

The deepest layers of the mouse brain cortex are L5 and L6, where L6 can further be split into L6a and L6b. L6b exhibits both a distinct location and gene expression profile from L6a, which tends to be closer to L5. Using scHybridNMF, we found that the seqFISH+ dataset showed clear location- and gene expression-based evidence for a distinct L6b excitatory neuron cell type. Tasic et al. (2016) give marker genes for L6a and 6b excitatory neurons, which are Rprm and Ctgf, respectively. In the seqFISH+ dataset, these exhibited strong spatial coherency, where we observed a clear boundary between cells that highly express Rprm vs Ctgf (**Supplementary Figure S4**), which clearly divided the two types of L6 excitatory neurons.

scHybridNMF was able to recover L6b excitatory neurons better than HMRF. To measure the differential gene expression across each cluster found by HMRF and scHybridNMF, we measured the expression of Ctgf and Cplx3, marker genes cited by Tasic et al. (2016), in **Figures 6A,B**. Because both genes were markers for L6b excitatory neurons, high-quality clusters are expected to exhibit a strongly distinct level of expression for these genes. We used the normalized, log-transformed gene expressions to create box plots of the expression statistics across each cluster. The side-by-side analysis of the two algorithms showed that the L6b cluster found by scHybridNMF exhibits a more distinct pattern of gene expression than the L6b cluster found by HMRF.

The region of cells highly expressing Ctgf in **Supplementary Figure S4** was small and sliver-like, and it bordered the rightmost side of layer 6. We found that the spatial location of the L6b cluster from scHybridNMF seemed to align more closely to this shape than the cluster from HMRF (**Supplementary Figure S5**). The cluster from HMRF included cells that were part of L6a.

3.3.3 scHybridNMF Refines Marker Gene Lists

Reducing False Positives of Layer 5 Excitatory Neuron Markers

The marker gene lists noted by Tasic et al. (2016) and by Dries et al. (2021) provided a basis for cell type annotations and interpretations of results in subsequent research. However, the markers obtained in Tasic et al. (2016) were based on scRNA-seq data only, and some of the location-specific marker genes may not actually demonstrate the expected location pattern. Indeed, from the DE analysis based on the clusters obtained by scHybridNMF, we found there were certain marker genes noted by Tasic et al.

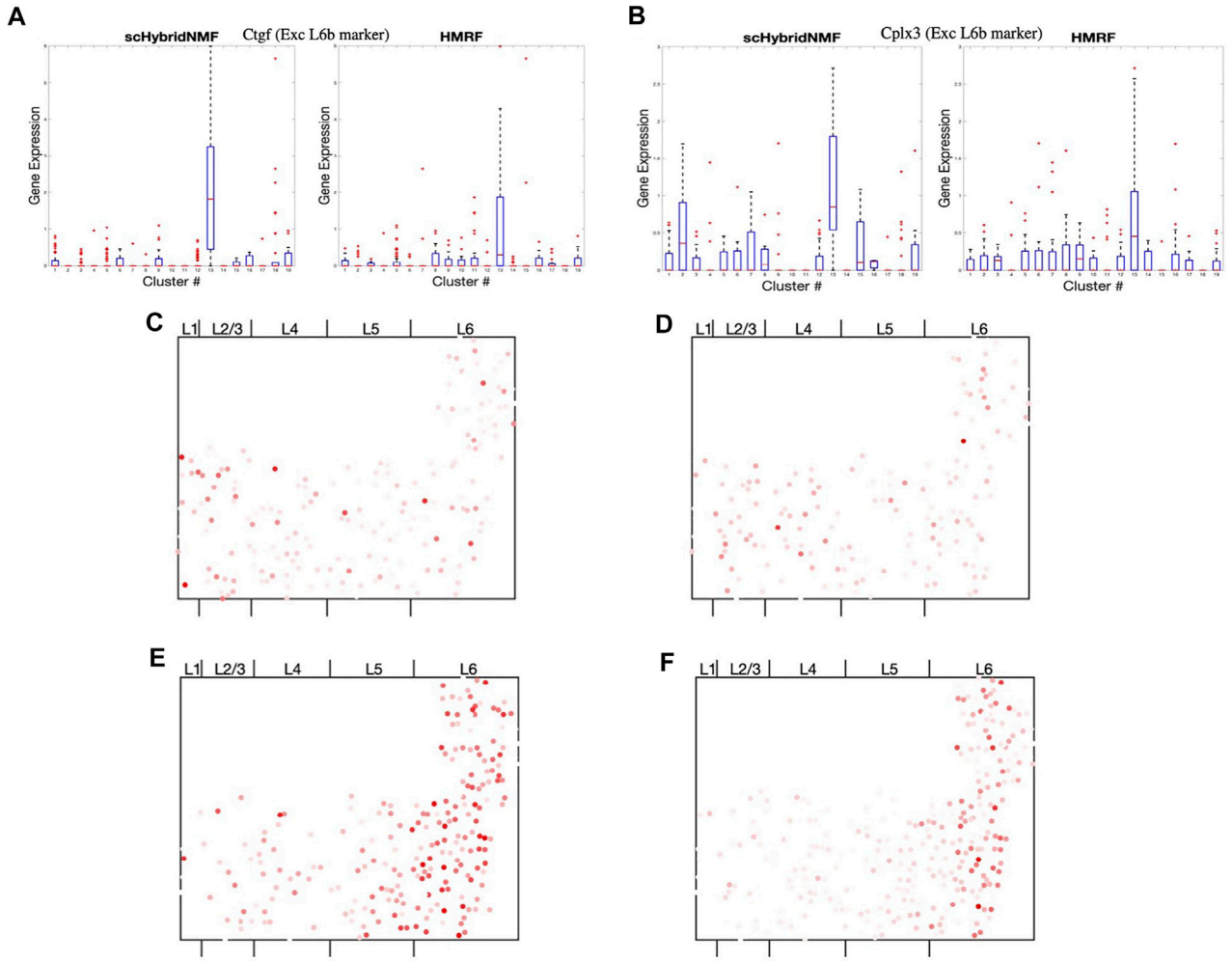


FIGURE 6 | (A,B) Box plots of the expressions of Ctgf and Cplx3 across each cell, grouped by cluster. **(C,D)** Dot plots of the expressions of Cpne2 and Ptgfr across each cell. These genes are marker genes of L5a and L5b excitatory neurons (Tasic et al., 2016). **(C)** Cpne2 (L5a marker). **(D)** Ptgfr (L5b marker). **(E,F)** Dot plots of the expressions of Islr2 and Rprm across each cell. We propose Islr2, a DE gene recovered through Scran on scHybridNMF clusters, as a marker gene for L6 excitatory neurons. We show the expression of Rprm as a baseline. **(E)** Islr2 (scHybridNMF). **(F)** Rprm (L6a, Tasic, *et al*).

(2016) that did not exist in the DE results. We focused on the marker genes for L5 excitatory neurons and further investigated the spatial pattern of these genes.

Tasic et al. (2016) catalogued 3 separate excitatory neuron types corresponding to L5. They were L5, L5a, and L5b excitatory neurons, where L5a and L5b distinguish the shallower and deeper regions of L5, respectively. The L5 excitatory neuron type referenced the entirety of layer 5. Of the 10,000 genes measured in seqFISH+, we found 17 were labeled as marker genes for only L5, L5a, or L5b excitatory neurons in Tasic et al. (2016). However, none of these genes exhibited any particular spatial pattern associated with L5. Examples of the spatial patterns are given in **Figures 6C,D** and **Supplementary Figure S6**.

Potential New Marker Gene for L6a Excitatory Neurons

Cluster 1 of scHybridNMF was annotated as L6a excitatory neurons both by gene expression and cell locations (**Supplementary Table S2**). Rprm is a marker gene from Tasic et al. (2016), and it exhibited a strong, spatially-conserved pattern in the seqFISH+ data (**Figure 6F**. We found another gene, Islr2, as a potential marker gene for L6a excitatory neurons. This is because it was differentially-expressed in cluster 1 [through Scran (Lun et al., 2016)], exhibited strong spatial cohesiveness, and was involved in neuron function and development (Abudureyimu et al., 2018) (**Figure 6E**). It was also found to be spatially concentrated in L5/6 by Giotto (Dries et al., 2021).

4 CONCLUSION AND DISCUSSION

We presented a hybrid clustering approach that can better identify cell types by incorporating sparse NMF and k-means clustering, which work well on high-dimensional gene expression and low-dimensional

location data. We demonstrated the robustness of scHybridNMF through experiments on both simulated and real data.

We showed that the hybrid framework was particularly useful when the performance of sparse NMF was affected by a low number of genes profiled or high within-cluster heterogeneity. scHybridNMF also outperformed k-means clustering under realistic scenarios. Through combining two classical methods for clustering, sparse NMF and k-means, scHybridNMF made better use of both data than either of the standalone methods as well as an existing method HMRF.

We also observed that scHybridNMF found biologically-meaningful clusters within real data. We analyzed the biological relevance of the clusters using cluster-specific DE genes that were found using cell cluster membership information. However, similar metagene analysis can be done using W_A, the cluster representative matrix. This matrix, which contains the final gene expression representatives of each cluster, was built using cell location and gene expression information. As such, W_A is constructed in such a way that incorporates both sources of information, and analyzing the differential expression of genes across different cluster representatives is intuitive. Each row of W_A corresponds to each gene, and the more variation of values there is in a row, the more likely the corresponding gene is biologically meaningful for cell type identification.

scHybridNMF is inherently flexible, owing to its matrix low-rank approximation formulation. As such, it can be extended via additional matrix terms and constraints to include more types of data or to perform biclustering. For example, we can include potential gene-gene interaction data to perform co-clustering of both cells and genes. The inferred gene clusters can be further used to study regulatory mechanisms in the cells and reconstruct gene regulatory networks.

DATA AVAILABILITY STATEMENT

scHybridNMF is available at github.com/soobleck/scHybridNMF. The simulated data and processed real data used in this study are also in the same GitHub repository. Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

HP and XZ conceived the study. All authors developed the methods. SO implemented the methods and drafted the manuscript. All authors edited and approved the manuscript.

FUNDING

This work was supported in part by NSF DBI-2019771. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

ACKNOWLEDGMENTS

We thank our colleagues for their editorial comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.763263/full#supplementary-material

REFERENCES

Abudureyimu, S., Asai, N., Enomoto, A., Weng, L., Kobayashi, H., Wang, X., et al. (2018). Essential Role of Linx/islr2 in the Development of the Forebrain Anterior Commissure. *Scientific Rep.* 8, 7292. doi:10.1038/s41598-018-24064-0

Bertsekas, D. P. (1997). Nonlinear Programming. *J. Oper. Res. Soc.* 48, 334. doi:10.1057/palgrave.jors.2600425

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., et al. (2015). Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. *Science* 348, 910–914. doi:10.1126/science.aab1601

Dries, R., Zhu, Q., Eng, C.-H. L., Li, H., Liu, K., Fu, Y., et al. (2021). Giotto: a Toolbox for Integrative Analysis and Visualization of Spatial Expression Data. *Genome Biol.* 22, 78. doi:10.1186/s13059-021-02286-2

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., et al. (2018). Integrative Analysis of Single-Cell Genomics Data by Coupled Nonnegative Matrix Factorizations. *Proc. Natl. Acad. Sci. USA* 115, 7723–7728. doi:10.1073/pnas.1805681115

Efremova, M., and Teichmann, S. A. (2020). Computational Methods for Single-Cell Omics across Modalities. *Nat. Methods* 17, 14–17. doi:10.1038/s41592-019-0692-4

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., et al. (2019). Transcriptome-scale Super-resolved Imaging in Tissues by RNA seqFISH+. *Nature* 568, 235–239. doi:10.1038/s41586-019-1049-y

Grippo, L., and Sciandrone, M. (2000). On the Convergence of the Block Nonlinear Gauss-Seidel Method under Convex Constraints. *Operations Res. Lett.* 26, 127–136. doi:10.1016/s0167-6377(99)00074-7

Jin, S., Zhang, L., and Nie, Q. (2020). scAI: an Unsupervised Approach for the Integrative Analysis of Parallel Single-Cell Transcriptomic and Epigenomic Profiles. *Genome Biol.* 21, 25. doi:10.1186/s13059-020-1932-8

Kim, H., and Park, H. (2007). Sparse Non-negative Matrix Factorizations via Alternating Non-negativity-constrained Least Squares for Microarray Data Analysis. *Bioinformatics* 23, 1495–1502. doi:10.1093/bioinformatics/btm134

Kim, J., He, Y., and Park, H. (2014). Algorithms for Nonnegative Matrix and Tensor Factorizations: a Unified View Based on Block Coordinate Descent Framework. *J. Glob. Optim* 58, 285–319. doi:10.1007/s10898-013-0035-4

Kim, J., and Park, H. (2008). "Toward Faster Nonnegative Matrix Factorization: A New Algorithm and Comparisons," in Proc. 8th IEEE ICDM 2008 (ICDM'08) (IEEE), 353–362. doi:10.1109/icdm.2008.149

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., et al. (2019). Identifying Gene Expression Programs of Cell-type Identity and Cellular Activity with Single-Cell RNA-Seq. *Elife* 8, e43803. doi:10.7554/eLife.43803

Kuang, D., Yun, S., and Park, H. (2015). SymNMF: Nonnegative Low-Rank Approximation of a Similarity Matrix for Graph Clustering. *J. Glob. Optim* 62, 545–574. doi:10.1007/s10898-014-0247-2

Lun, A., McCarthy, D., and Marioni, J. (2016). A Step-by-step Workflow for Low-Level Analysis of Single-Cell Rna-Seq Data with Bioconductor. *F1000Res* 5, 2122. doi:10.12688/f1000research.9501.2

Mayr, U., Serra, D., and Liberali, P. (2019). Exploring Single Cells in Space and Time during Tissue Development, Homeostasis and Regeneration. *Development* 146, dev176727. doi:10.1242/dev.176727

McKinley, K. L., Castillo-Azofeifa, D., and Klein, O. D. (2020). Tools and Concepts for Interrogating and Defining Cellular Identity. *Cell Stem Cell* 26, 632–656. doi:10.1016/j.stem.2020.03.015

Morris, S. A. (2019). The Evolving Concept of Cell Identity in the Single Cell Era. *Development* 146, dev169748. doi:10.1242/dev.169748

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., et al. (2017). Multiplexed Quantification of Proteins and Transcripts in Single Cells. *Nat. Biotechnol.* 35, 936–939. doi:10.1038/nbt.3973

Shao, C., and Höfer, T. (2017). Robust Classification of Single-Cell Transcriptome Data by Nonnegative Matrix Factorization. *Bioinformatics* 33, 235–242. doi:10.1093/bioinformatics/btw607

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics. *Science* 353, 78–82. doi:10.1126/science.aaf2403

Stuart, T., and Satija, R. (2019). Integrative Single-Cell Analysis. *Nat. Rev. Genet.* 20, 257–272. doi:10.1038/s41576-019-0093-7

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., et al. (2016). Adult Mouse Cortical Cell Taxonomy Revealed by Single Cell Transcriptomics. *Nat. Neurosci.* 19, 335–346. doi:10.1038/nn.4216

Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., et al. (2018). Three-dimensional Intact-Tissue Sequencing of Single-Cell Transcriptional States. *Science* 361. doi:10.1126/science.aat5691

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-Cell Multi-Omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* 177, 1873–1887. doi:10.1016/j.cell.2019.05.006

Zhang, X., Xu, C., and Yosef, N. (2019). Simulating Multiple Faceted Variability in Single Cell RNA Sequencing. *Nat. Commun.* 10, 2611. doi:10.1038/s41467-019-10500-w

Zhu, Q., Shah, S., Dries, R., Cai, L., and Yuan, G.-C. (2018). Identification of Spatially Associated Subpopulations by Combining scRNAseq and Sequential Fluorescence *In Situ* Hybridization Data. *Nat. Biotechnol.* 36, 1183–1190. doi:10.1038/nbt.4260

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 November 2021
doi: 10.3389/fgene.2021.760155

Inferring Differential Networks by Integrating Gene Expression Data With Additional Knowledge

Chen Liu[1], *Dehan Cai*[2], *WuCha Zeng*[1] *and Yun Huang*[3]*

[1]Department of Chemotherapy, The First Affiliated Hospital of Fujian Medical University, Fuzhou, China, [2]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China, [3]Department of Geriatric Medicine, The First Affiliated Hospital of Fujian Medical University, Fuzhou, China

Evidences increasingly indicate the involvement of gene network rewiring in disease development and cell differentiation. With the accumulation of high-throughput gene expression data, it is now possible to infer the changes of gene networks between two different states or cell types via computational approaches. However, the distribution diversity of multi-platform gene expression data and the sparseness and high noise rate of single-cell RNA sequencing (scRNA-seq) data raise new challenges for existing differential network estimation methods. Furthermore, most existing methods are purely rely on gene expression data, and ignore the additional information provided by various existing biological knowledge. In this study, to address these challenges, we propose a general framework, named weighted joint sparse penalized D-trace model (WJSDM), to infer differential gene networks by integrating multi-platform gene expression data and multiple prior biological knowledge. Firstly, a non-paranormal graphical model is employed to tackle gene expression data with missing values. Then we propose a weighted group bridge penalty to integrate multi-platform gene expression data and various existing biological knowledge. Experiment results on synthetic data demonstrate the effectiveness of our method in inferring differential networks. We apply our method to the gene expression data of ovarian cancer and the scRNA-seq data of circulating tumor cells of prostate cancer, and infer the differential network associated with platinum resistance of ovarian cancer and anti-androgen resistance of prostate cancer. By analyzing the estimated differential networks, we find some important biological insights about the mechanisms underlying platinum resistance of ovarian cancer and anti-androgen resistance of prostate cancer.

Keywords: single-cell RNA sequencing, differential network analysis, prior information, graphical model, gene regulatory network

OPEN ACCESS

Edited by:
Min Wu,
*Institute for Infocomm Research (A*STAR), Singapore*

Reviewed by:
Mengwei Li,
*Singapore Immunology Network (A*STAR), Singapore*
Wen Zhang,
Huazhong Agricultural University, China

***Correspondence:**
Yun Huang
huangyun20150318@163.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *17 August 2021*
Accepted: *13 October 2021*
Published: *11 November 2021*

Citation:
Liu C, Cai D, Zeng W and Huang Y (2021) Inferring Differential Networks by Integrating Gene Expression Data With Additional Knowledge. Front. Genet. 12:760155. doi: 10.3389/fgene.2021.760155

1 INTRODUCTION

Biological systems often involve the complex regulatory relationships between genes, which could change substantially in different states or developmental stages. Inferring the changes of gene regulatory networks between two different states or cell types is important for revealing the regulatory mechanisms relevant to disease development and cell differentiation (Tian et al., 2016; Zhang et al., 2017). With the accumulation of state-specific gene expression data, a great number of computational approaches have been proposed for estimating gene regulatory networks as well as their difference between two distinct states from gene expression data (Danaher et al., 2014;

Ha et al., 2015; Lichtblau et al., 2016; Tian et al., 2016; Zhang et al., 2016; Ou-Yang et al., 2017; Uppal et al., 2018).

Due to the ability in capturing the conditional dependencies among genes, Gaussian graphical models have been widely used to infer gene regulatory networks (Danaher et al., 2014; Zhang et al., 2016; Yuan et al., 2017; Ou-Yang et al., 2019). Existing Gaussian graphical model-based differential network estimation methods can be roughly divided into two categories, i.e., indirect estimation models (Danaher et al., 2014; Zhang et al., 2016) and direct estimation models (Tian et al., 2016; Yuan et al., 2017). Indirect estimation models first estimate each state-specific network separately and then infer the differential network by calculating the difference between two state-specific networks (Danaher et al., 2014). Whereas direct estimation models directly estimate the difference between two state-specific networks without the need to estimate individual state-specific networks (Tian et al., 2016). As the number of parameters that needs to be estimated in direct estimation models is half of that in indirect estimation models, direct estimation models usually achieve better performance than indirect estimation models in differential network estimation, especially in the case of small sample size (Yuan et al., 2017).

Although the above models have been successfully used to infer differential networks (Danaher et al., 2014; Tian et al., 2016; Yuan et al., 2017), they are mainly designed for bulk tissue gene expression data collected from a single data platform. Recently, with the development of high-throughput experimental technologies, we are able to collect bulk gene expression data of same samples from multiple data platforms. As the gene expression data collected from different data platforms may provide some shared and specific information about the regulatory relationships between genes, integrating multi-platform gene expression data could help to improve the accuracy of differential network estimation (Zhang et al., 2016, 2017). Moreover, the advance of single-cell RNA sequencing (scRNA-seq) techniques offers a great opportunity for inferring the regulatory relationships between genes at single cell resolution. The accumulation of scRNA-seq data paves the way to infer cell-type-specific gene networks, which could help to explore the heterogeneity between different cell types (Pratapa et al., 2020). However, due to technical limitations of existing scRNA-seq technologies, a truly expressed gene may not be identified in some cells, which leads to excess of false zeros in scRNA-seq data (i.e., dropout events) (Stegle et al., 2015). Existing differential network estimation models usually assume that the observed data are complete, and rarely consider missing value problem. To handle the distribution diversity of multi-platform gene expression data and the sparseness of single-cell RNA sequencing (scRNA-seq) data, Ou-Yang et al. (2021) proposed an indirect differential network estimation model, which can integrate the gene expression data collected from multiple data platforms and tackle the missing value problem. Moreover, their model can take into account the changes in gene expression levels when inferring differential networks.

The above models only use gene expression data to infer differential networks. However, since the number of samples are usually much smaller than the number of genes, and scRNA-seq data are much sparser and noisier than bulk RNA-seq data, it is difficult to infer differential networks accurately only based on gene expression data. Besides gene expression data, existing knowledge of genes and knowledge of the regulatory relationships among genes may also help to improve the accuracy of differential network estimation (Xu et al., 2018). For example, we can collect some literature-curated gene regulatory interactions from public database (Han et al., 2015). As the changes of regulatory relationships between two different states is more likely to occur between genes that are known to have regulatory interactions, considering prior gene regulatory interactions may help to improve the accuracy of differential network estimation. Moreover, researchers have found that genes within same pathways usually interact with each other to carry out their biological functions, and genes belong to different pathways seldom interact with each other (Wu et al., 2019). Thus, taking into account pathway information may also facilitate the inference of differential networks.

In this study, to address the above problems and provide a differential network estimation method that can generally work well on different types of data, we propose a novel method named Weighted Joint Sparse penalized D-trace Model (WJSDM). Our model can directly estimate the differential networks between two different states by integrating multi-platform gene expression data with additional biological knowledge. Similar to (Ou-Yang et al., 2021), based on non-paranormal graphical model and revised Kendall's tau correlation, our model can tackle non-Gaussian data with missing values, which make it able to deal with multi-platform gene expression and scRNA-seq data. By using D-trace loss function, our model can estimate the differential network directly, which reduce the number of parameters that need to be estimated. To integrate various prior biological knowledge and take into account changes in gene expression levels, we propose a weighted group bridge penalty. Our model can be solved by using an accelerated proximal gradient method. Simulation studies are first conducted to evaluate the performance of our model. According to the experiment results, our model can always achieve better performance than other state-of-the-art differential network estimation models, which demonstrate the effectiveness of our model in integrating prior information and handling gene expression data with missing values. Extensive experiments on two real data sets also demonstrate the advantages of our model in inferring differential networks and revealing the underlying mechanisms of disease developments. The source code of our proposed model is available at https://github.com/Yunhuang85/WJSDM.

2 METHODS

In this section, we will first review the non-paranormal distribution and D-trace loss. Then we will introduce our weighted joint sparse penalized D-trace model.

2.1 Non-paranormal Distribution

Let $X = (X_1, X_2, \ldots, X_p)$ denote a p-dimensional random vector which follows a multivariate normal distribution $X \sim N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ is the covariance matrix. For multivariate normal distributions, X_i is independent of X_j given the other variables if and only if the corresponding entry in the inverse covariance matrix (precision matrix) $\Theta = \Sigma^{-1}$ is equal to zero, i.e., $\Theta_{ij} = 0$. Thus, the conditional dependence relationships among p random variables in X can be obtained by identifying the nonzero elements in Θ. However, the normal distribution assumption is too restrictive in practice. To relax the normal distribution assumption, non-paranormal distribution is proposed. $X = (X_1, X_2, \ldots, X_p)$ is said to follow a non-paranormal distribution $X \sim NPN(f, \Sigma)$ if there exists a set of monotone and differentiable functions $\{f_j\}_{j=1}^{p}$ such that $f(X) = (f_1(X_1), \ldots, f_p(X_p)) \sim N(0, \Sigma)$. It has been proven that $\Theta = \Sigma^{-1}$ encodes the conditional dependence relationships among X. That is, X_i is independent of X_j given the other variables if and only if $\Theta_{ij} = 0$.

2.2 D-Trace Loss

Given the gene expression data $\{X^{(c)}\}_{c=1,2}$ of two different states. Each data set $X^{(c)} \in \mathbb{R}^{n_c\times p}$ includes n_c samples and p common genes. Suppose the n_c samples within each data set are from the same non-paranormal distribution $NPN(f^{(c)}, \Sigma^{(c)})$, where $\Sigma^{(c)} \in \mathbb{R}^{p\times p}$ is the covariance matrix. The conditional dependence relationships between these p genes can be inferred from the precision matrix $\Theta^{(c)} = (\Sigma^{(c)})^{-1}$. Thus, the difference between two state-specific networks can be presented as $\Delta = \Theta^{(2)} - \Theta^{(1)}$. To estimate the differential network Δ efficiently, we can utilize the following D-trace loss function (Yuan et al., 2017), which could directly estimate the difference between two precision matrices without separate estimation of each precision matrix:

$$\underset{\Delta=\Delta^T}{arg\ min}\ L_D\left(\Delta; \Sigma^{(1)}, \Sigma^{(2)}\right) = \frac{1}{4}\left(\langle \Sigma^{(1)}\Delta, \Delta\Sigma^{(2)}\rangle + \langle \Sigma^{(2)}\Delta, \Delta\Sigma^{(1)}\rangle\right) - \langle \Delta, \Sigma^{(1)} - \Sigma^{(2)}\rangle. \tag{1}$$

where $(A, B) = tr(AB^T)$. In practice, we need to use the sample covariance matrices $\hat{\Sigma}^{(c)}$ and minimize $L_D(\Delta; \hat{\Sigma}^{(1)}, \hat{\Sigma}^{(2)})$ with respect to Δ to calculate the estimator of Δ. For non-paranormal distribution, the sample non-paranormal covariance matrix can be computed *via* rank-based correlation estimator, e.g., Kendall's tau correlation, without estimating the univariate transformation functions $f^{(c)}$.

2.3 Notations and Problem Statement

Assuming that there are two different groups of samples. As the gene expression data of same samples can be collected from multiple data platforms, suppose we can observe the expression levels of p common genes for these two groups of samples from K different data platforms, and the cth group contains n_c samples, $c = 1, 2$. Let $X^{(kc)} \in \mathbb{R}^{n_c\times p}$ denote the gene expression matrix of the cth group collected from kth platform, where n_c and p denote the number of samples and the number of common genes, respectively. Suppose the n_c samples are from the same non-paranormal distribution $NPN(f^{(kc)}, \Sigma^{(kc)})$, where $\Sigma^{(kc)} \in \mathbb{R}^{p\times p}$ is the covariance matrix. Let $\{\Theta^{(kc)}\}_{k=1,\ldots,K}^{c=1,2}$ denote the precision matrices for two groups of samples collected from K platforms, where $\Theta^{(kc)} = (\Sigma^{(kc)})^{-1}$. For samples collected from the kth platform, the difference between two state-specific networks can be presented as $\Delta^{(k)} = \Theta^{(k2)} - \Theta^{(k1)}$. Our goal is to estimate K differential networks $\{\Delta^{(k)}\}_{k=1,\ldots,K}$ jointly. For the sake of convenience, we denote $\{X^{(kc)}\}_{k=1,\ldots,K}^{c=1,2}$, $\{\Sigma^{(kc)}\}_{k=1,\ldots,K}^{c=1,2}$, $\{\Theta^{(kc)}\}_{k=1,\ldots,K}^{c=1,2}$ and $\{\Delta^{(k)}\}_{k=1,\ldots,K}$ as $\{X^{(kc)}\}$, $\{\Sigma^{(kc)}\}$, $\{\Theta^{(kc)}\}$ and $\{\Delta^{(k)}\}$, respectively.

2.4 Weighted Joint Sparse Penalized D-Trace Model

The above D-trace loss is designed to infer the differential network between two different groups of samples from a single data platform, and cannot utilize the common information provided by gene expression data collected from multiple data platforms. Thus, in this study, we extend D-trace loss and develop a weighted joint sparse D-trace model (WJSDM), which can draw support from gene expression data collected from multiple data platforms to estimate the differential network between two different states.

According to the above D-trace loss, the loss function L_{KD} of K data platforms can be given by:

$$L_{KD}\left(\{\Delta^{(k)}\}\right) = \frac{1}{4}\sum_{k=1}^{K}\left(\langle \hat{\Sigma}^{(k1)}\Delta^{(k)}, \Delta^{(k)}\hat{\Sigma}^{(k2)}\rangle + \langle \hat{\Sigma}^{(k2)}\Delta^{(k)}, \Delta^{(k)}\hat{\Sigma}^{(k1)}\rangle\right) - \sum_{k=1}^{K}\left(\langle \Delta^{(k)}, \hat{\Sigma}^{(k1)} - \hat{\Sigma}^{(k2)}\rangle\right). \tag{2}$$

where $\hat{\Sigma}^{(kc)}$ is the sample non-paranormal covariance matrix for cth group and kth data platform, $k = 1, \ldots, K$ and $c = 1, 2$. As gene expression data may include some missing values, similar to (Wang et al., 2014; Ou-Yang et al., 2021), we adopt a rank-based correlation, i.e., revised Kendall's tau correlation, to estimate $\hat{\Sigma}^{(kc)}$. In particular, let $n_{ij}^{(kc)} = \sum_{1\le l\le n_{kc}} d_{li}^{(kc)} d_{lj}^{(kc)}$ denote the number of samples in the cth group and kth platform that have nonzero expression values for ith and jth genes simultaneously, where $d_{lj}^{(kc)} = 1$ if $X_{lj}^{(kc)} \neq 0$ and $d_{lj}^{(kc)} = 0$ if $X_{lj}^{(kc)} = 0$. The revised Kendall's tau correlation between ith and jth genes are defined as follows:

$$\hat{\tau}_{ij}^{(kc)} = \frac{1}{n_{ij}^{(kc)}\left(n_{ij}^{(kc)} - 1\right)} \sum_{l\neq l'} d_{li}^{(kc)} d_{l'i}^{(kc)} d_{lj}^{(kc)} d_{l'j}^{(kc)} sign\left(\left(X_{li}^{(kc)} - X_{l'i}^{(kc)}\right)\left(X_{lj}^{(kc)} - X_{l'j}^{(kc)}\right)\right) \tag{3}$$

As Kendall's tau correlation are invariant under strictly monotone marginal transformations (Liu et al., 2012), $\Sigma_{ij}^{(kc)}$ can be estimated by the following definition of $\hat{\Sigma}_{ij}^{(kc)}$

$$\hat{\Sigma}_{ij}^{(kc)} = \begin{cases} sin\left(\frac{\pi}{2}\hat{\tau}_{ij}^{(kc)}\right), & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases} \tag{4}$$

In this study, each sample non-paranormal covariance matrix $\hat{\Sigma}_{ij}^{(kc)}$ is computed according to the revised Kendall's tau correlation and transformation function defined in **Eqs 3**, **4**. To ensure $\hat{\Sigma}^{(kc)}$ is positive semidefinite, following Zhang et al. (2021) and Higham (1988), we compute the nearest positive semidefinite matrix $S^{(kc)}$ of $\hat{\Sigma}^{(kc)}$ and use it to replace $\hat{\Sigma}^{(kc)}$.

Note that the differential networks inferred from gene expression data collected from different data platforms may share certain network structures, and the differential networks between two different states may be sparse. Furthermore, differentially expressed genes usually tend to change their regulatory relationships with other genes. Thus, to jointly estimate multiple differential networks and consider the changes in expression levels of individual genes when inferring differential networks, similar to (Ou-Yang et al., 2021), we introduce the following group bridge penalty function:

$$P\left(\left\{\Delta^{(k)}\right\}\right) = \sum_{i,j} \sqrt{\sum_{k=1}^{K} \tau_{ij}^{(k)} |\Delta_{ij}^{(k)}|}. \tag{5}$$

where $\tau_{ij}^{(k)} = 1 - (1 - r_i^{(k)})(1 - r_j^{(k)})$ can assign different weights to different pairs of genes, and $r_i^{(k)} \in [0, 1]$ denotes the parameter which measures the differential expression level of ith gene, inferred from the kth experimental platform. In this study, following Ou-Yang et al. (2021), the p-value of Wilcoxon rank-sum test is used to calculate $r_i^{(k)}$, which can reflect the differential expression level of ith gene. With this penalty function, the differential networks $\{\Delta^{(k)}\}$ inferred from K different data platforms may have similar patterns of sparsity and have some shared and specific edges.

Besides gene expression data, there are usually some prior biological knowledge that can help to improve the accuracy of differential network estimation, such as pathway information and prior gene interactions. To incorporate these prior information when inferring differential networks, we extend the above group bridge penalty function to the following weighted group bridge penalty function:

$$P\left(\left\{\Delta^{(k)}\right\}\right) = \sum_{i,j} W_{ij} \sqrt{\sum_{k=1}^{K} \tau_{ij}^{(k)} |\Delta_{ij}^{(k)}|}. \tag{6}$$

Here, $W = [W_{ij}]$ is the weight matrix defined by prior knowledge. In this study, the prior information we used includes pathway information and gene interactions that have been verified from other biological studies. Let $G \in \{0,1\}^{p\times p}$ and $F \in \{0,1\}^{p\times p}$ denote the prior gene interaction and co-pathway indication matrices, respectively, where $G_{ij} = 1$ if the ith and jth genes are known to have regulatory relationship and $G_{ij} = 0$ otherwise, $F_{ij} = 1$ if the ith and jth genes belong to at least one common pathway and $F_{ij} = 0$ otherwise. To assign different weights to different pairs of genes, we define W_{ij} as follows:

$$W_{ij} = \begin{cases} w_g, & \text{if } G_{ij} = 1, \\ 1, & \text{if } G_{ij} = 0 \text{ and } F_{ij} = 1, \\ w_f, & \text{if } G_{ij} = 0 \text{ and } F_{ij} = 0. \end{cases} \tag{7}$$

where w_g and w_f are two predefined weight parameters. In reality, the differential edges are more likely to take place between gene pairs that are known to have interactions, and the differential edges are less likely to occur between genes that belong to different pathways. Thus, to assign small penalties to gene pairs that are known to have interactions and large penalties to gene pairs that belong to different pathways, the value of w_g should be less than 1 and the value of w_f should be larger than 1. Following previous studies (Xu et al., 2018), in this study, we fix $w_g = 0.3$ and $w_f = 10$.

By combining (2) and (6), the objective function of our Weighted Joint Sparse penalized D-trace Model (WJSDM) is given by:

$$\left\{\hat{\Delta}^{(k)}\right\} = \underset{\left\{\Delta^{(k)} = \left(\Delta^{(k)}\right)^T\right\}}{arg\ min} \; L_{KD}\left(\left\{\Delta^{(k)}\right\}\right) + \lambda \sum_{i,j} W_{ij} \sqrt{\sum_{k=1}^{K} \tau_{ij}^{(k)} |\Delta_{ij}^{(k)}|}. \tag{8}$$

where λ is a non-negative tuning parameter to control the sparsity levels of the estimated differential networks. We use an iterative approach based on local linear approximation (Zou and Li, 2008) and the accelerated proximal gradient method (Parikh and Boyd, 2014; Xu et al., 2018) to solve problem (**Eq. 8**).

According to (Yuan et al., 2017), the gradient of the D-trace loss function with respect to Δ takes the following form:

$$\nabla L_D(\Delta) = \frac{1}{2}\left(\hat{\Sigma}^{(1)}\Delta\hat{\Sigma}^{(2)} + \hat{\Sigma}^{(2)}\Delta\hat{\Sigma}^{(1)}\right) - \left(\hat{\Sigma}^{(1)} - \hat{\Sigma}^{(2)}\right). \tag{9}$$

Following the proximal gradient method (Parikh and Boyd, 2014), L_{KD} can be approximated by the following function:

$$\tilde{L}_{KD}\left(\left\{\Delta^{(k)}\right\}; \left\{\left(\hat{\Delta}^{(k)}\right)^{(t)}\right\}, \{l_k\}\right) = \sum_{k=1}^{K}\Big[L_D\left(\left(\hat{\Delta}^{(k)}\right)^{(t)}\right) + tr\left(\nabla L_D\left(\left(\hat{\Delta}^{(k)}\right)^{(t)}\right)\left(\Delta^{(k)} - \left(\hat{\Delta}^{(k)}\right)^{(t)}\right)\right) + \frac{1}{2l_k}\|\Delta^{(k)} - \left(\hat{\Delta}^{(k)}\right)^{(t)}\|_F^2\Big]. \tag{10}$$

where $(\hat{\Delta}^{(k)})^{(t)}$ is the estimation of $\Delta^{(k)}$ at tth iteration, $l_k > 0$ and $\|A\|_F^2 = \sum_{i,j=1}^{p} A_{ij}$. We rewrite the $\tilde{L}_{KD}$ function as:

$$\tilde{L}_{KD}\left(\left\{\Delta^{(k)}\right\}; \left\{\left(\hat{\Delta}^{(k)}\right)^{(t)}\right\}, \{l_k\}\right) = \sum_{k=1}^{K}\Big[\frac{1}{2l_k}\|\Delta^{(k)} - \left(\left(\hat{\Delta}^{(k)}\right)^{(t)} - l_k\nabla L_D\left(\left(\hat{\Delta}^{(k)}\right)^{(t)}\right)\right)\|_F^2 + \varphi\left(\left(\hat{\Delta}^{(k)}\right)^{(t)}\right)\Big]. \tag{11}$$

where $\varphi((\hat{\Delta}^{(k)})^{(t)})$ is a function of $(\hat{\Delta}^{(k)})^{(t)}$.

Algorithm 1. Complete Algorithm for WJSDM (8)

- **Input:** $2K$ matrices $\hat{\Sigma}^{(kc)}$, prior gene interaction network G, co-pathway indication matrix F, and tuning parameters λ.
- **Output:** Estimated K differential networks $\{\hat{\Delta}^{(k)}\}$.
- **Main algorithm:**

1) Initialize $\hat{\Delta}^{(k)}$ and $l_k = 1$ for $k = 1, \ldots, K$.
2) Compute the weight matrix W.
3) **While** not converged **do**
4) $\quad \phi_{ij} = \frac{\tau_{ij}^{(k)}}{2\sqrt{\sum_{k=1}^{K} \tau_{ij}^{(k)} |(\hat{\Delta}_{ij}^{(k)})^{(t)}|}}$.
5) $\quad$ **for** $k = 1 : K$
6) $\quad\quad B_k^{(t+1)} = (\hat{\Delta}^{(k)})^{(t)} + \frac{t}{t+3}((\hat{\Delta}^{(k)})^{(t)} - (\hat{\Delta}^{(k)})^{(t-1)})$.
7) $\quad\quad$ **Repeat**
8) $\quad\quad\quad Z = soft(B_k^{(t+1)} - l_k^{(t)} \bigtriangledown L_D(B_k^{(t+1)}), l_k^{(t)}\lambda W \circ \Phi)$.
9) $\quad\quad\quad$ **break if** $L_D(Z) \leqslant \tilde{L}_D(B_k^{(t+1)}; Z, l_k^{(t)})$.
10) $\quad\quad\quad$ Update $l_k^{(t)} = \frac{1}{2} l_k^{(t)}$;
11) $\quad\quad$ **Return** $(\hat{\Delta}^{(k)})^{(t+1)} = Z, l_k^{(t+1)} = l_k^{(t)}$.
12) $\quad$ **End for**
13) $\quad t = t + 1$;
14) $\quad$ Check the convergence condition.
15) **End While**
16) **Return** $\left\{\hat{\Delta}^{(k)}\right\} = \left\{(\hat{\Delta}^{(k)})^{(t)}\right\}$.

Annotation:
Convergence condition: $|F(\{(\hat{\Delta}^{(k)})^{(t+1)}\}) - F(\{(\hat{\Delta}^{(k)})^{(t)}\})| \leqslant \varepsilon |F(\{(\hat{\Delta}^{(k)})^{(t+1)}\})|$, where $\varepsilon = 10^{-5}$.

According to local linear approximation (Parikh and Boyd, 2014), **Eq. 6** can be approximated as:

$$P\left(\left\{\Delta^{(k)}\right\}\right) \approx \lambda \sum_{k=1}^{K} \sum_{i,j} \phi_{ij} W_{ij} |\Delta_{ij}^{(k)}|. \tag{12}$$

where $\phi_{ij} = \frac{\tau_{ij}^{(k)}}{2\sqrt{\sum_{k=1}^{K} \tau_{ij}^{(k)} |(\hat{\Delta}_{ij}^{(k)})^{(t)}|}}$. Therefore, at $(t + 1)$-th iteration, problem (**Eq. 8**) can be decomposed into the following K individual optimization problems:

$$\begin{aligned}\left(\hat{\Delta}^{(k)}\right)^{(t+1)} = \underset{\Delta^{(k)}=\left(\Delta^{(k)}\right)^T}{arg\ min}\ & \frac{1}{2}\|\Delta^{(k)} \\ & - \left(\left(\hat{\Delta}^{(k)}\right)^{(t)} - l_k \nabla L_D\left(\left(\hat{\Delta}^{(k)}\right)^{(t)}\right) \right)\|_F^2 \\ & + \lambda l_k \sum_{i,j} \phi_{ij} W_{ij} |\Delta_{ij}^{(k)}|. \end{aligned} \tag{13}$$

The solution of our WJSDM is summarized in **Algorithm 1**. The computational complexity of each iteration in **Algorithm 1** is $O(Kp^3 + Kp^2)$, where K is the number of data platforms and p is the number of genes.

2.5 Parameter Selection

There is a tuning parameter λ in WJSDM, which affects the sparsity level of the estimated differential networks. In this study, following previous studies (Zhang et al., 2016), we use a stability approach, named StARS method (Liu et al., 2010; Meinshausen and Bühlmann, 2010), to determine the value of λ. The detailed procedure of our parameter selection method is summarized in **Algorithm 2**.

Algorithm 2. Tuning Parameter Selection for WJSDM

Step1: Randomly generate S sample subsets D_1,...,D_S from each sample set, the number of samples in each subset accounts for 80% of the total number of samples.

Step2: Set the value range Γ for parameter λ.

Step3: Main selection step:

(1) Estimate K differential networks $\left\{\hat{\Delta}_s^{(k)}(\lambda)\right\}_{k=1,\ldots,K}$ for each subsets D_s and each λ from Γ. According to the method of StARS, we select the optimal value of λ by solving the following problem:

$$\lambda_{opt} = arg \min_{\xi \in \Gamma} \left\{ \max_{\lambda \geq \xi} \sum_{k=1}^{K} 2(Stab^k)/K \leq \sigma \right\} \tag{14}$$

where $Stab^k = \sum_{i<j} \bar{a}_{i,j}^k (1 - \bar{a}_{i,j}^k) / \binom{p}{2}$, $\bar{a}_{i,j}^k = \frac{1}{S} \sum_{s=1}^{S} I(\ (i,j) \in \hat{\Delta}_s^{(k)}(\lambda)\ )$, $I(\cdot)$ is an indicator function.

3 RESULTS

In this section, we first perform simulation studies to assess the performance of our proposed WJSDM. Then we apply our model on real data sets.

3.1 Simulation Studies

To demonstrate the effectiveness of our WJSDM in inferring differential networks, we compare WJSDM with five state-of-the-art differential network estimation models, i.e., FGL (Danaher et al., 2014), TDJGL (Zhang et al., 2016), WDNE (Ou-Yang et al., 2021), GGL (Danaher et al., 2014) and D-trace (Yuan et al., 2017).

3.1.1 Data Generation

In this simulation study, we consider two groups of samples and their observations on p common genes collected from $K = 3$ data platforms, and generate six scale-free networks for the two groups of samples and three data platforms. Here, we set $p = 100$ and generate $n_1 = n_2 = n \in \{50, 100, 200\}$ observations for each data platform. Each network includes three pathways with $0.4p$ genes per pathway, and there are $0.2p$ genes shared by the second and third pathway. For each pathway, we choose 10% edges as differential edges. To model the heterogeneity between different data platforms, we choose 10% of differential edges to be platform-specific. Since differentially expressed genes tend to change their regulatory relationships with other genes, we select 30% genes as differentially expressed genes and the edges connected to differentially expressed genes are more likely to be differential edges. There are no differential edges between genes belong to different pathways. To make a fair comparison with Gaussian graphical model-based methods, the gene expression levels of each cell are simulated by using a multivariate normal distribution. To generate the prior gene interaction network G, we select a prior rate δ of nonzero elements from the above six scale-free networks randomly and connect the corresponding genes in G. Note that gene expression data may include missing values. In this study, the expression values of a gene may be lost randomly, and the missing rate is set to $\tau \in \{0.6, 0.8\}$.

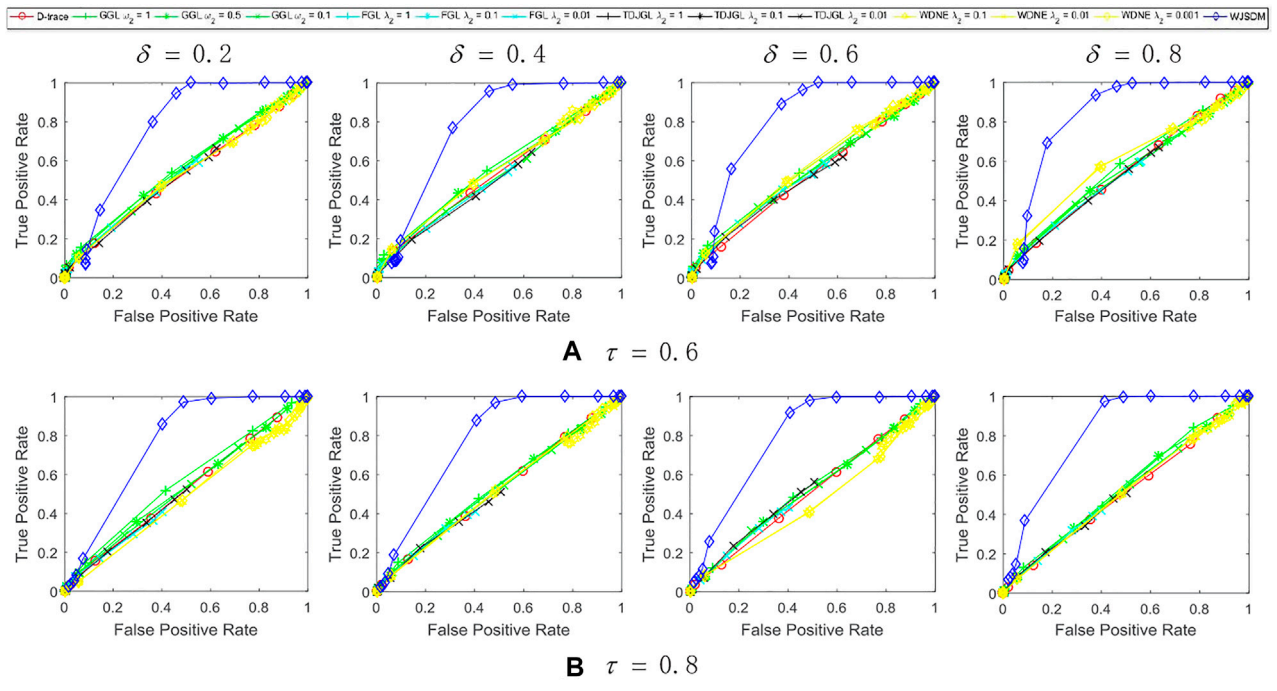


FIGURE 1 | Performance of WJSDM, D-trace, GGL, FGL, TDJGL and WDNE with $p = 100$, $K = 3$, $n = 50$ and missing rate **(A)** $\tau = 0.6$, **(B)** $\tau = 0.8$. Within each plot, each line presents the performance of a method with the value of λ (for WJSDM and D-trace), λ_1 (for FGL, TDJGL and WDNE) or ω_1 (for GGL) varying for a fixed value of λ_2 (for FGL, TDJGL and WDNE) or ω_2 (for GGL). Each curve is obtained by averaging the performance of a method over ten random generations of data.

3.1.2 Simulation Results

Let $\hat{\Delta}^{(k)}$ (for indirect estimation methods: $\hat{\Delta}^{(k)} = \hat{\Theta}^{(k1)} - \hat{\Theta}^{(k2)}$) denote the estimated differential network between two states for the kth platform, and $\Delta^{(k)}$ denote the real differential network for the kth platform. We use the following two metrics to evaluate the performance of various algorithms:

$$TPR = \frac{\sum_{k=1}^{K}\sum_{i<j} I\left(\hat{\Delta}_{ij}^{(k)} \neq 0 \quad \text{and} \quad \Delta_{ij}^{(k)} \neq 0\right)}{\sum_{k=1}^{K}\sum_{i<j} I\left(\Delta_{ij}^{(k)} \neq 0\right)},$$

$$FPR = \frac{\sum_{k=1}^{K}\sum_{i<j} I\left(\hat{\Delta}_{ij}^{(k)} \neq 0 \quad \text{and} \quad \Delta_{ij}^{(k)} = 0\right)}{\sum_{k=1}^{K}\sum_{i<j} I\left(\Delta_{ij}^{(k)} = 0\right)}.$$

where TPR denotes true positive rate, FPR denotes false positive rate, and $I(\cdot)$ is an indicator function.

As all methods have some hyper-parameters that need to be predefined, we generate a series of solutions for each model with different values of hyper-parameters, and assess their performances. In particular, for FGL, GGL, TDJGL and WDNE, there are two parameters λ_1 and λ_2. While for D-trace and our WJSDM, there is one parameter λ. To ease interpretation, following Danaher et al. (2014), the tuning parameters for GGL are reparameterized as $\omega_1 = \lambda_1 + \frac{1}{\sqrt{2}}\lambda_2$ and $\omega_2 = \frac{1}{\sqrt{2}}\lambda_2/(\lambda_1 + \frac{1}{\sqrt{2}}\lambda_2)$. The experiment results of all methods are averaged over 10 random generations of synthetic data. **Figures 1–3** show the performance of various methods on synthetic data. The columns of each figure show the results of various methods with different values of prior rate δ, and the rows of this figure show the results with different values of missing rate τ. In this figure, each plot shows the $TPR - FPR$ curves of various methods. Within each plot, different colored lines present the performances of different methods and different points in each line indicate the results with respect to different values of hyper-parameters. The colored lines for D-trace and WJSDM indicate their results as the values of λ varied. The colored lines for FGL, GGL, TDJGL and WDNE are obtained by fixing the value of λ_2 (or ω_2 for GGL) and varying the values of λ_1 (or ω_1 for GGL). For λ_1 and ω_1, we choose 15 values ranging from 0.01 to 10 (for WDNE, the value of λ_1 is ranging from 0.01 to 100).

We can find from these figures that our WJSDM outperforms other compared methods in all cases. GGL can estimate multiple networks that share common network structures, but it cannot identify the differences between different networks. FGL and D-trace can infer the changes between different networks, but they cannot integrate the data collected from different data platforms. TDJGL is an extension of FGL, which can integrate multi-platform gene expression data. WDNE is an extension of TDJGL, which can handle gene expression data with missing values and take into account changes in gene expression levels. All of the above methods cannot make use of the prior information provided by additional knowledge when inferring differential networks. WDNE is a indirect differential network estimation model, which need to estimate the state-specific networks in advance. Thus, when the sample size is small, it cannot estimate differential network accurately. As shown in **Figure 3**, when the sample size is large, WDNE can achieve good performance and outperform other compared methods in most cases. The superior

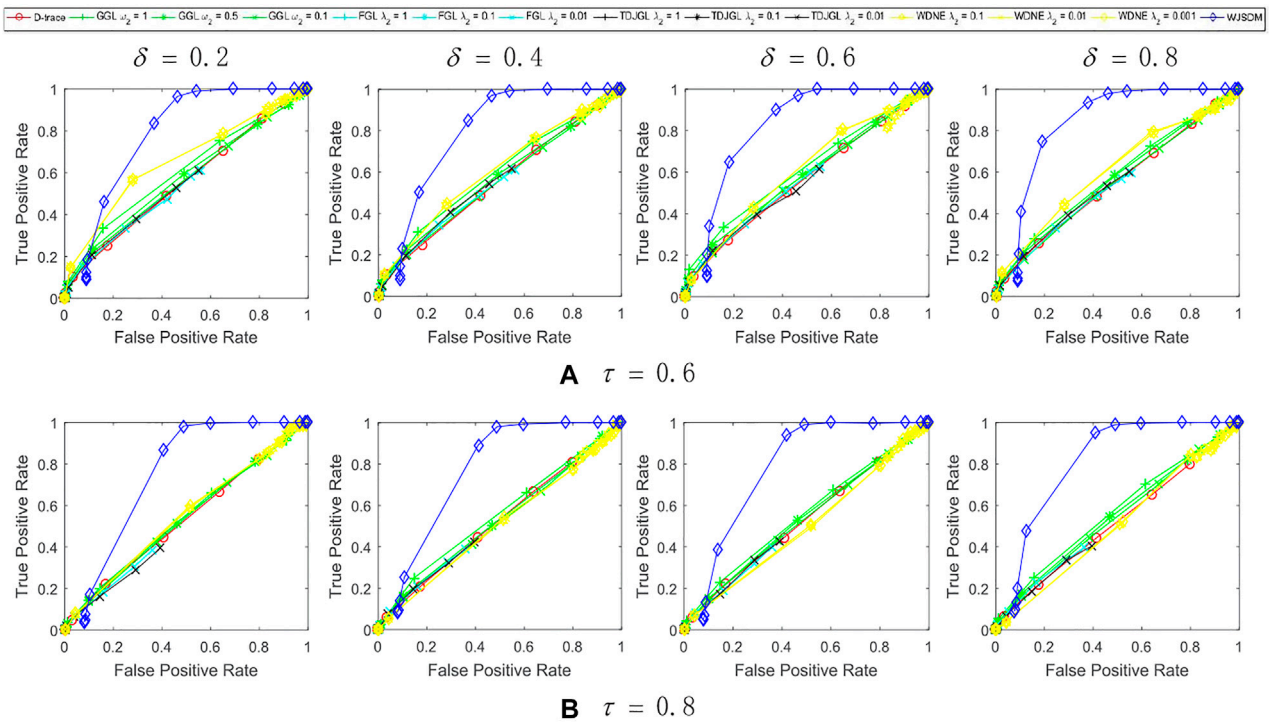


FIGURE 2 | Performance of WJSDM, D-trace, GGL, FGL, TDJGL and WDNE with $p = 100$, $K = 3$, $n = 100$ and missing rate **(A)** $\tau = 0.6$, **(B)** $\tau = 0.8$. Within each plot, each line presents the performance of a method with the value of λ (for WJSDM and D-trace), λ_1 (for FGL, TDJGL and WDNE) or ω_1 (for GGL) varying for a fixed value of λ_2 (for FGL, TDJGL and WDNE) or ω_2 (for GGL). Each curve is obtained by averaging the performance of a method over ten random generations of data.

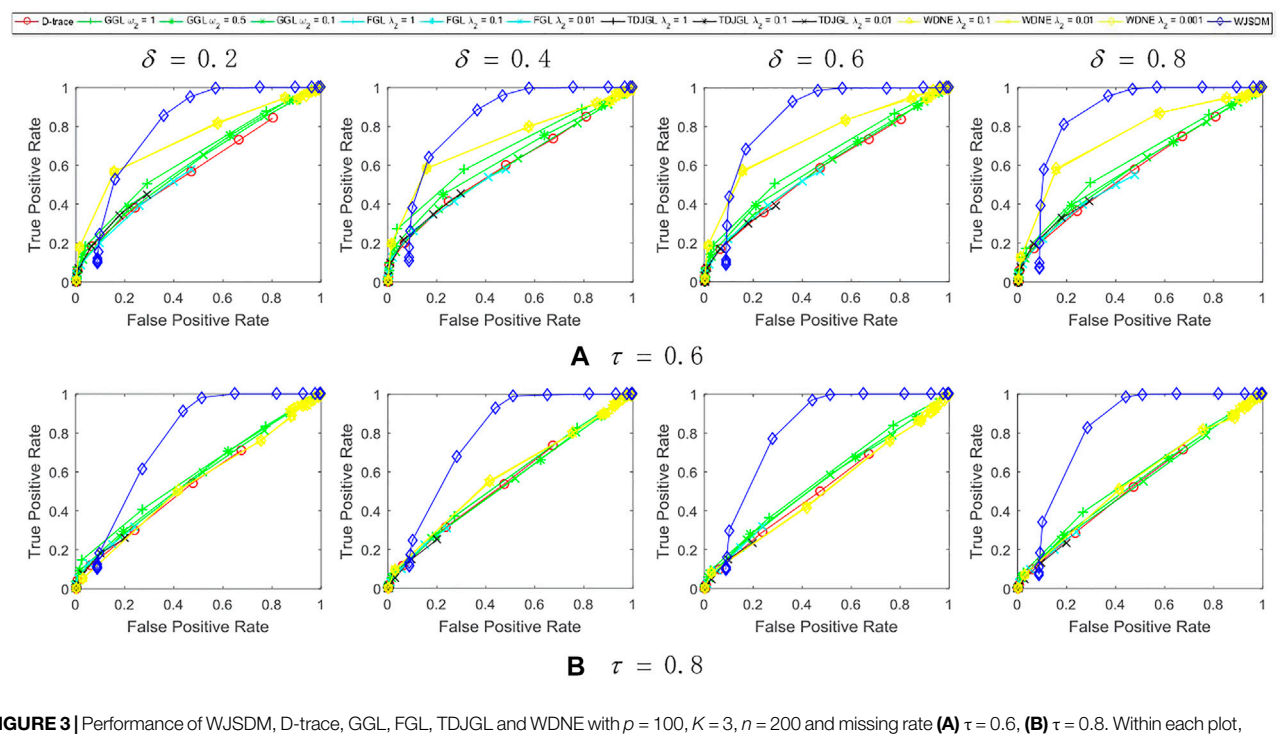


FIGURE 3 | Performance of WJSDM, D-trace, GGL, FGL, TDJGL and WDNE with $p = 100$, $K = 3$, $n = 200$ and missing rate **(A)** $\tau = 0.6$, **(B)** $\tau = 0.8$. Within each plot, each line presents the performance of a method with the value of λ (for WJSDM and D-trace), λ_1 (for FGL, TDJGL and WDNE) or ω_1 (for GGL) varying for a fixed value of λ_2 (for FGL, TDJGL and WDNE) or ω_2 (for GGL). Each curve is obtained by averaging the performance of a method over ten random generations of data.

performance of WJSDM over WDNE demonstrates the benefit of inferring differential network directly and integrating multiple additional knowledge.

3.2 Real Data Analysis

3.2.1 Ovarian Cancer Analysis

Platinum agents, represented by cisplatin, are the most active cytotoxic drugs in ovarian cancer (Tapia and Diaz-Padilla, 2013). Women with platinum-resistant ovarian cancer continue to have poor survival rates, and effective treatment of platinum resistance still remains the largest unmet need in ovarian cancer (van Zyl et al., 2018). To explore the underlying mechanisms of platinum resistance, we utilize WJSDM to infer the changes of gene regulatory networks between platinum-sensitive and platinum-resistant ovarian tumors. In particular, we collect three gene expression datasets from TCGA database (Network, 2011), which measure gene expression levels from three platforms, i.e., Agilent 244K Custom Gene Expression G450, Affymetrix HT Human Genome U133 Array Plate Set and Affymetrix Human Exon 1.0 ST Array. The expression levels of 8,417 genes for 512 samples are available for all these three platforms. Among the 512 samples, 97 samples are platinum-resistant and 243 samples are platinum-sensitive. Following Zhang et al. (2017), we focus our analysis on seven critical pathways involved in platinum resistance, i.e., apoptosis, cell cycle, ErbB signaling pathway, mismatch repair, nucleotide excision repair, p53 signaling pathway and platinum drug resistance (Kanehisa and Goto, 2000). There are 315 genes in our datasets that belong to these seven pathways. The prior gene interaction network is downloaded from the TRRUST database (Han et al., 2015). There are 361 prior interactions among the 315 genes.

According to the parameter selection strategy (i.e., StARS) introduced above, the tuning parameter λ of our WJSDM is set to $\lambda = 2.5$. The estimated differential network between platinum-resistant and platinum-sensitive tumors, which describes the changes of gene regulatory relationships associated with platinum resistance, is shown in **Figure 4**. Since we are not able to obtain the true differential network between platinum-resistant and platinum-sensitive tumors, it is hard to measure the accuracy of the estimated differential networks. In fact, a common challenge in evaluating the performance of differential network estimation on real data sets is the lack of reference data. Hub genes in the differential network have more differential edges, which means they may play more important roles in driving the resistance of platinum. Thus, in this study, following previous studies (Zhang et al., 2016, 2017; Ou-Yang et al., 2019), we investigate the functions of the hub genes in our estimated differential network. In particular, the top 10 genes with the highest degree in our estimated differential network are considered as hubs (**Table 1**). To verify whether our identified hub genes are related to platinum resistance in ovarian cancer, similar to (Zhang et al., 2017), we draw support from six public datasets. In particular, we collect 161 cisplatin resistance-related genes and 758 drug resistance-related genes from the database of Genomic Elements Associated with drug Resistance (GEAR) (Wang et al., 2017), 548 experimentally verified ovarian cancer-related genes from the ovarian cancer gene database (OCGene) (Liu et al., 2015), 116 anti-cancer drug targets from the cancer drug resistance database (CancerDR) (Kumar et al., 2013), 572 cancer genes from the Cancer Gene Census database (Futreal et al., 2004) and 3,545 regulator genes from (Grechkin et al., 2016). Among the identified 10 hub genes, five of them are cisplatin resistance-related genes, eight of them are drug resistance-related genes, six of them are ovarian cancer-related genes, five of them are anti-cancer drug targets, four of them are cancer genes and nine of them are regulator genes.

Note that the above six public datasets are still far from complete. Thus, we also draw support from literature search to explore whether our identified hub genes are related to cisplatin resistance in ovarian cancer. Among these genes, BBC3 has been reported to be associated with cisplatin resistance in ovarian cancer (Zhang et al., 2012; Grozav et al., 2015), and has been proposed as a chemosensitizer in platinum compounds-based ovarian cancer therapy (Yuan et al., 2011). PARP1 have been shown to involved in cisplatin resistance in ovarian cancer, and could be treated as a potential sensitizer in cisplatin chemotherapy (Liu et al., 2018). TP73 has been found to be associated with clinical responsiveness to platinum-based chemotherapy in advanced non-small cell lung cancer (NSCLC) (Yuan et al., 2006). Researches have found that TP73 could be a genetic marker for ovarian response (Bakay et al., 2021). Thus, it is interesting to study the association between TP73 and platinum resistance in ovarian cancer.

We can also find from **Table 1** that our identified hub genes include both differentially (in this study, genes whose p-values are less than 0.05 are treated as differentially expressed genes) and non-differentially expressed genes. For example, MAPK8, CCND1, TP53, CDKN1A and BCL2 are related to cisplatin resistance in ovarian cancers. None of these five genes showed differential expression between platinum-resistant and platinum-sensitive tumors. Thus, our model can identify functional important genes that cannot be found by differential expression analysis, which demonstrate the superiority of our model over differential expression analysis.

3.2.2 Prostate Cancer Analysis

Enzalutamide is a second-generation anti-androgen medication which has been used in the treatment of prostate cancer (Scher et al., 2012). However, the mechanisms underlying the resistance of enzalutamide remain vague. We then apply WJSDM to the scRNA-seq data of circulating tumor cells of prostate cancer, and investigate the changes of gene regulatory relationships that associated with enzalutamide-resistant. In particular, we collect a scRNA-seq data set of prostate circulating tumor cells from GEO database with accession number: GSE67980 (Miyamoto et al., 2015). There are 77 samples isolated from 13 patients, where 41 samples are enzalutamide-naive and 36 samples are enzalutamide-resistant (Chiu et al., 2018). Among 21,696 genes, 7,508 genes have no sequencing reads in all the 77 samples. We focus our analysis on three critical pathways download from the Kyoto Encyclopedia of Genes and Genomes database (Kanehisa and Goto, 2000), i.e., Notch signaling pathway, Wnt signaling pathway and PI3K-AKT signaling pathway. By removing genes

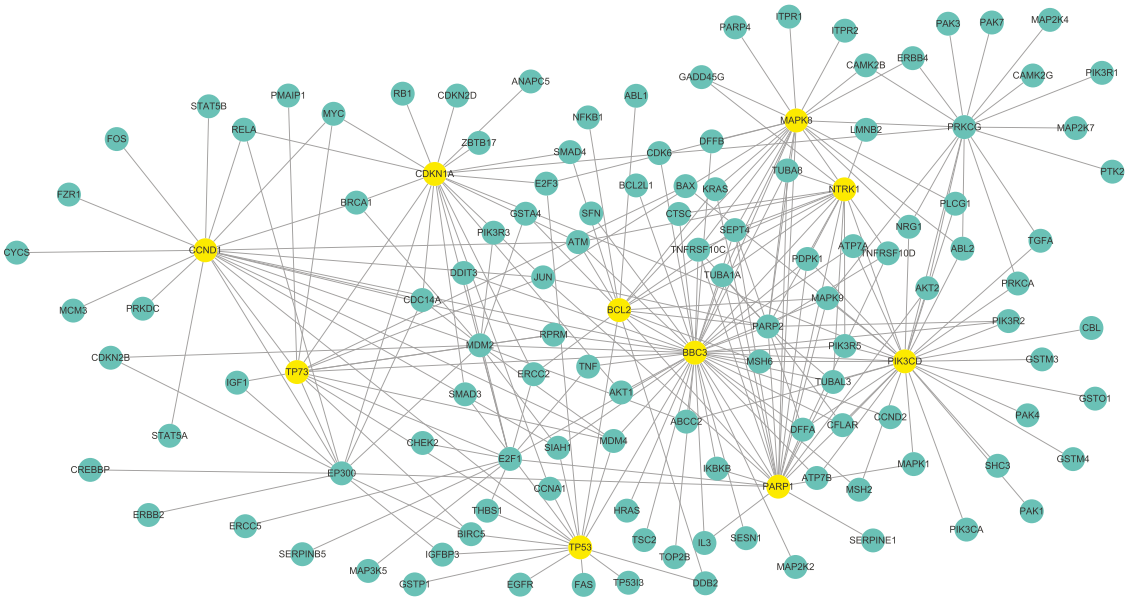


FIGURE 4 | The differential network between platinum-resistant and platinum-sensitive tumors identified by WJSDM. Here, yellow nodes denote the top-10 hub genes in the differential network.

TABLE 1 | Top-10 hub genes in the estimated differential gene network between platinum-resistant and platinum-sensitive tumors.

Rank	Gene	Degree	*p*-value	CR	DR	OCG	ADT	CG	RG
1	BBC3	9|38|35	0.023|0.016|0.024						√
2	MAPK8	9|21|23	0.306|0.853|0.495	√	√	√	√		√
3	PIK3CD	7|23|21	0.005|0.006|0.001				√		
4	PARP1	7|19|20	0.058|0.028|0.003		√	√	√		√
5	CCND1	8|10|21	0.125|0.462|0.071	√	√	√		√	√
6	TP53	4|16|16	0.702|0.681|0.957	√	√	√		√	√
7	CDKN1A	4|10|21	0.519|0.557|0.146	√	√	√			√
8	TP73	9|11|15	0.073|0.854|0.270		√				√
9	BCL2	5|13|16	0.592|0.493|0.167	√	√	√	√	√	√
10	NTRK1	11|16|4	0.011|0.098|0.945		√		√	√	√

If a gene is a cisplatin resistance-related gene (CR), drug resistance-related gene (DR), ovarian cancer gene (OCG), anti-cancer drug target (ADT), cancer gene (CG) or regulator gene (RG), there is an √ in the corresponding entry. a|b|c§ *represents the degree and* p*-values (computed by Wilcoxon rank-sum test) of genes in the differential networks inferred from three platforms, respectively.*

with no sequencing reads, there are 234 genes in the scRNA-seq data that belong to these three pathways. The prior gene interaction network is downloaded from the TRRUST database (Han et al., 2015). There are 178 prior interactions among the 234 genes.

According to the parameter selection strategy (i.e., StARS) introduced above, the tuning parameter λ of our WJSDM is set to $\lambda = 0.7197$. The estimated differential network between enzalutamide-resistant and enzalutamide-naive samples, which describes the changes of gene regulatory relationships associated with enzalutamide resistance, is shown in **Figure 5**. Hub genes in the differential network have more differential edges, which means they may play more important roles in driving the resistance of enzalutamide. Thus, we investigate the functions of the hub genes in our estimated differential network. In particular, the top 10 genes with the highest degree in our estimated differential network are considered as hubs (as shown in **Table 2**). We can find from **Table 2** that all of these 10 hub genes are related to prostate cancer and five of them are associated with enzalutamide-resistant.

Among these genes, MYC has been reported to be implicated in the development of enzalutamide resistance and the increase of MYC expression is correlated with shorter progression free survival in patients undergoing enzalutamide treatment (Handle et al., 2019; Furlan et al., 2021). However, this gene does not show differential expression between enzalutamide-resistant and enzalutamide-naive samples. Thus, it cannot be found by differential expression analysis. RAC1, which has been demonstrated to be upregulated in enzalutamide-resistant prostate cancer cells, plays a crucial role in enzalutamide resistance and could be a potential target for the treatment of

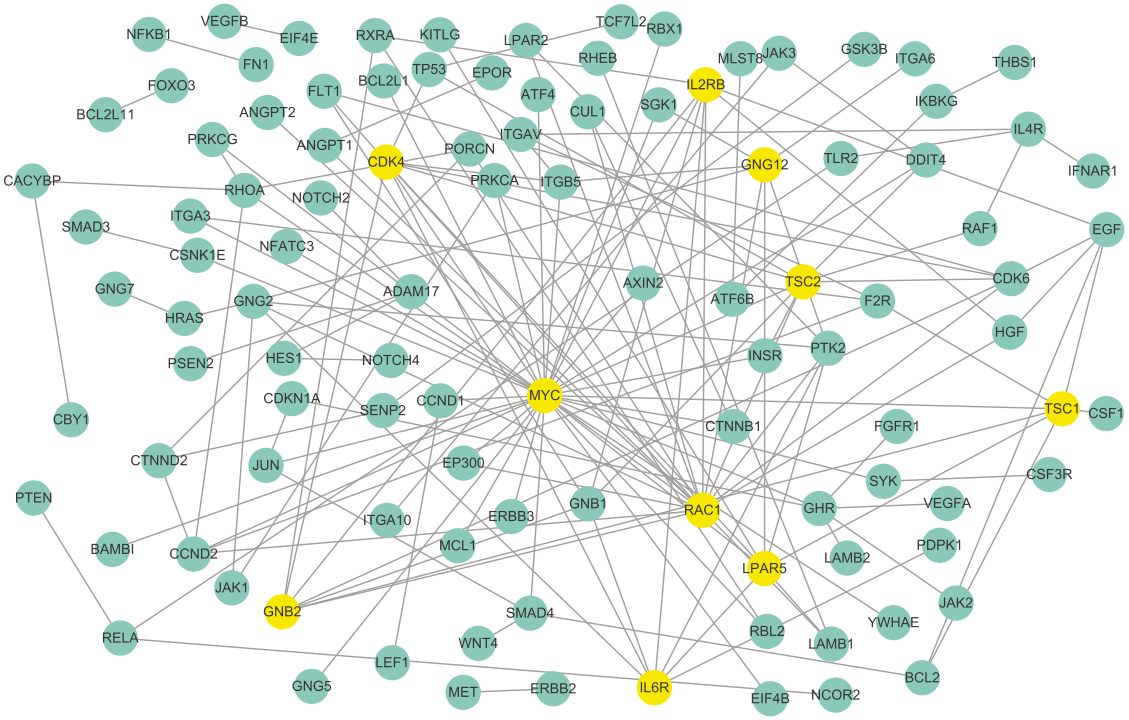


FIGURE 5 | The differential network between enzalutamide-resistant and enzalutamide-naive samples identified by WJSDM. Here, yellow nodes denote the top-10 hub genes in the differential network.

TABLE 2 | Top-10 hub genes in the estimated differential gene network between enzalutamide-resistant and enzalutamide-naive samples.

Rank	Gene	Degree	*p*-value	PCa	ER
1	MYC	35	0.15	○	○
2	RAC1	18	5.60e-5	○	○
3	CDK4	10	7.84e-4	○	○
4	TSC2	9	0.005	○	○
5	IL2RB	8	0.005	○	
6	LPAR5	8	0.027	○	
7	GNB2	7	0.010	○	
8	GNG12	7	0.024	○	
9	IL6R	7	0.031	○	○
10	TSC1	7	0.010	○	

If the gene is associated with prostate cancer (PCa) or enzalutamide-resistant (ER) according to literature search (Wu et al., 2006; Tam et al., 2007; Lin et al., 2015; Wang et al., 2018; Handle et al., 2019; Chen et al., 2020; Kase et al., 2020; Balijepalli et al., 2021; Dickson et al., 2021; Furlan et al., 2021), a ○ is placed in the corresponding entry. The p-value of each gene is computed by Wilcoxon rank-sum test.

castration-resistant prostate cancer (Chen et al., 2020). Knockdown of TSC1 and TSC2 have been shown to promote the proliferation of prostate cancer cells Lin et al. (2015). LPAR5 has been reported to be involved with immune response inhibition and cancer progression Geraldo et al. (2021). Researches have found that GRB2 is associated with shorter survival of patients with aggressive prostate cancer (Iwata et al., 2021). The activation of the IL-6R/JAK/STAT3 pathway has been found to be involved with the development of hormonerefractory prostate cancer (Tam et al., 2007). The combined inhibition of IL6R and HMGB1 has been reported to be a new treatment for enzalutamide resistance in patients with advanced prostate cancer (Wang et al., 2018).

The above results demonstrate the effectiveness of our WJSDM in inferring the difference between the gene networks of different disease states, and provide important insights about the underlying regulatory mechanisms of the platinum resistance in ovarian cancer and the enzalutamide resistance in prostate cancer.

4 CONCLUSION

Increasing evidences indicate the changes of gene regulatory relationships between different cell states or developmental stages, which motivate the development of computational models to infer differential networks. In this paper, based on gene expression data and additional biological knowledge, we propose a novel differential network estimation method named weighted joint sparse penalized D-trace model (WJSDM), to infer the changes of gene regulatory networks between two different states. By employing D-trace loss function and using a revised Kendall's tau correlation, our method can directly infer the differential network between two different states from gene expression data with missing

values. Furthermore, to integrate the gene expression data collected from different data platforms and utilize the information provided by various prior biological knowledge, we propose a weighted group bridge penalty function, which enable our model to draw support from multiple related data sets. Experiment results on synthetic data sets show that compared with other state-of-the-art differential network estimation methods, our method can infer differential networks more accurately. We also apply our method to the gene expression data of ovarian tumors and circulating tumor cells of prostate cancer, and estimate the differential network associated with platinum resistance of ovarian cancer and anti-androgen resistance of prostate cancer. By analyzing our estimated differential networks, we find some important biological insights about the mechanisms underlying platinum resistance of ovarian cancer and anti-androgen resistance of prostate cancer.

With the development of single-cell sequencing techniques, an increasing number of single-cell multi-omics data are becoming available. How to efficiently integrate single-cell multi-omics data is an interesting future work. We will try to extend our model to handle this problem.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/ https://www.ncbi.nlm.nih.gov/geo/ https://www.kegg.jp/ https://www.grnpedia.org/trrust/.

AUTHOR CONTRIBUTIONS

CL and DC conceived and designed the study, performed the statistical analysis and drafted the manuscript. YH and WZ conceived of the study, and participated in its design and coordination and helped to draft and revise the manuscript. All authors read and approved the final manuscript.

FUNDING

This work is supported by the Natural Science Foundation of Fujian Province, China (Grant No. 2020J01956) and Startup Fund for scientific research, Fujian Medical University (Grant No.2019QH1065, 2020QH1041).

REFERENCES

Bakay, K., Coban, U., Arslan, M. A., Guven, D., and Tural, S. (2021). Effects of Hrg and Tp73 Gene Variations on Ovarian Response. *Gynecol. Endocrinol.* 1, 1–5. doi:10.1080/09513590.2021.1974379

Balijepalli, P., Sitton, C. C., and Meier, K. E. (2021). Lysophosphatidic Acid Signaling in Cancer Cells: What Makes Lpa So Special. *Cells* 10, 2059. doi:10.3390/cells10082059

Chen, X., Yin, L., Qiao, G., Li, Y., Li, B., Bai, Y., et al. (2020). Inhibition of Rac1 Reverses Enzalutamide Resistance in Castration-resistant P-rostate C-ancer. *Oncol. Lett.* 20, 2997–3005. doi:10.3892/ol.2020.11823

Chiu, Y. C., Hsiao, T. H., Wang, L. J., Chen, Y., and Shao, Y. J. (2018). Scdnet: a Computational Tool for Single-Cell Differential Network Analysis. *BMC Syst. Biol.* 12, 124–166. doi:10.1186/s12918-018-0652-0

Danaher, P., Wang, P., and Witten, D. M. (2014). The Joint Graphical Lasso for Inverse Covariance Estimation across Multiple Classes. *J. R. Stat. Soc. B* 76, 373–397. doi:10.1111/rssb.12033

Dickson, M. A., Ravi, V., Ganjoo, K. N., and Iyer, G. (2021). Institutional Experience with Nab-Sirolimus in Patients with Malignancies Harboring Tsc1 or Tsc2 Mutations. *Jco* 39, 3111. doi:10.1200/jco.2021.39.15_suppl.3111

Furlan, T., Kirchmair, A., Sampson, N., Puhr, M., Gruber, M., Trajanoski, Z., et al. (2021). Myc-mediated Ribosomal Gene Expression Sensitizes Enzalutamide-Resistant Prostate Cancer Cells to Ep300/crebbp Inhibitors. *Am. J. Pathol.* 191, 1094–1107. doi:10.1016/j.ajpath.2021.02.017

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A Census of Human Cancer Genes. *Nat. Rev. Cancer* 4, 177–183. doi:10.1038/nrc1299

Geraldo, L. H. M., de Sampaio Spohr, T. C. L., do Amaral, R. F., da Fonseca, A. C. C., Garcia, C., de Almeida Mendes, F., et al. (2021). Role of Lysophosphatidic Acid and its Receptors in Health and Disease: Novel Therapeutic Strategies. *Signal. Transduction Targeted Ther.* 6, 1–18. doi:10.1038/s41392-020-00367-5

Grechkin, M., Logsdon, B. A., Gentles, A. J., and Lee, S.-I. (2016). Identifying Network Perturbation in Cancer. *Plos Comput. Biol.* 12, e1004888. doi:10.1371/journal.pcbi.1004888

Grozav, A., Balacescu, O., Balacescu, L., Cheminel, T., Berindan-Neagoe, I., and Therrien, B. (2015). Synthesis, Anticancer Activity, and Genome Profiling of Thiazolo Arene Ruthenium Complexes. *J. Med. Chem.* 58, 8475–8490. doi:10.1021/acs.jmedchem.5b00855

Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2015). Dingo: Differential Network Analysis in Genomics. *Bioinformatics* 31, 3413–3420. doi:10.1093/bioinformatics/btv406

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., et al. (2015). Trrust: a Reference Database of Human Transcriptional Regulatory Interactions. *Sci. Rep.* 5, 11432. doi:10.1038/srep11432

Handle, F., Prekovic, S., Helsen, C., Van den Broeck, T., Smeets, E., Moris, L., et al. (2019). Drivers of Ar Indifferent Anti-androgen Resistance in Prostate Cancer Cells. *Sci. Rep.* 9, 13786. doi:10.1038/s41598-019-50220-1

Higham, N. J. (1988). Computing a Nearest Symmetric Positive Semidefinite Matrix. *Linear algebra its Appl.* 103, 103–118. doi:10.1016/0024-3795(88)90223-6

Iwata, T., Sedukhina, A. S., Kubota, M., Oonuma, S., Maeda, I., Yoshiike, M., et al. (2021). A New Bioinformatics Approach Identifies Overexpression of Grb2 as a Poor Prognostic Biomarker for Prostate Cancer. *Sci. Rep.* 11, 5696–5698. doi:10.1038/s41598-021-85086-9

Kanehisa, M., and Goto, S. (2000). Kegg: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Kase, A. M., Copland, J. A., III, and Tan, W. (2020). Novel Therapeutic Strategies for Cdk4/6 Inhibitors in Metastatic Castrate-Resistant Prostate Cancer. *Ott* 13, 10499–10513. doi:10.2147/ott.s266085

Kumar, R., Chaudhary, K., Gupta, S., Singh, H., Kumar, S., Gautam, A., et al. (2013). Cancerdr: Cancer Drug Resistance Database. *Sci. Rep.* 3, 1445–1446. doi:10.1038/srep01445

Lichtblau, Y., Zimmermann, K., Haldemann, B., Lenze, D., Hummel, M., and Leser, U. (2016). Comparative Assessment of Differential Network Analysis Methods. *Brief Bioinform* 18, 837–850. doi:10.1093/bib/bbw061

Lin, H.-P., Lin, C.-Y., Huo, C., Jan, Y.-J., Tseng, J.-C., Jiang, S. S., et al. (2015). AKT3 Promotes Prostate Cancer Proliferation Cells through Regulation of Akt, B-Raf & TSC1/TSC2. *Oncotarget* 6, 27097–27112. doi:10.18632/oncotarget.4553

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability Approach to Regularization Selection (Stars) for High Dimensional Graphical Models. *Adv. Neural Inf. Process. Syst.* 24, 1432–1440.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional Semiparametric Gaussian Copula Graphical Models. *Ann. Stat.* 40, 2293–2326. doi:10.1214/12-aos1037

Liu, Q., Zhu, D., Hao, B., Zhang, Z., and Tian, Y. (2018). Luteolin Promotes the Sensitivity of Cisplatin in Ovarian Cancer by Decreasing Prpa1-Medicated Autophagy. *Cel Mol Biol (Noisy-le-grand)* 64, 17–22. doi:10.14715/cmb/2018.64.6.4

Liu, Y., Xia, J., Sun, J., and Zhao, M. (2015). Ocgene: a Database of Experimentally Verified Ovarian Cancer-Related Genes with Precomputed Regulation Information. *Cell Death Dis.* 6, e2036. doi:10.1038/cddis.2015.380

Meinshausen, N., and Bühlmann, P. (2010). Stability Selection. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 72, 417–473. doi:10.1111/j.1467-9868.2010.00740.x

Miyamoto, D. T., Zheng, Y., Wittner, B. S., Lee, R. J., Zhu, H., Broderick, K. T., et al. (2015). Rna-seq of Single Prostate Ctcs Implicates Noncanonical Wnt Signaling in Antiandrogen Resistance. *Science* 349, 1351–1356. doi:10.1126/science.aab0917

Network, T. C. G. A. R. (2011). Integrated Genomic Analyses of Ovarian Carcinoma. *Nature* 474, 609–615. doi:10.1038/nature10166

Ou-Yang, L., Cai, D., Zhang, X. F., and Yan, H. (2021). Wdne: an Integrative Graphical Model for Inferring Differential Networks from Multi-Platform Gene Expression Data with Missing Values. *Brief Bioinform*, bbab086. doi:10.1093/bib/bbab086

Ou-Yang, L., Yan, H., and Zhang, X.-F. (2017). Identifying Differential Networks Based on Multi-Platform Gene Expression Data. *Mol. Biosyst.* 13, 183–192. doi:10.1039/c6mb00619a

Ou-Yang, L., Zhang, X.-F., Zhao, X.-M., Wang, D. D., Wang, F. L., Lei, B., et al. (2019). Joint Learning of Multiple Differential Networks with Latent Variables. *IEEE Trans. Cybern.* 49, 3494–3506. doi:10.1109/tcyb.2018.2845838

Parikh, N., and Boyd, S. (2014). Proximal Algorithms. *FNT in Optimization* 1, 127–239. doi:10.1561/2400000003

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking Algorithms for Gene Regulatory Network Inference from Single-Cell Transcriptomic Data. *Nat. Methods* 17, 147–154. doi:10.1038/s41592-019-0690-6

Scher, H. I., Fizazi, K., Saad, F., Taplin, M.-E., Sternberg, C. N., Miller, K., et al. (2012). Increased Survival with Enzalutamide in Prostate Cancer after Chemotherapy. *N. Engl. J. Med.* 367, 1187–1197. doi:10.1056/nejmoa1207506

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and Analytical Challenges in Single-Cell Transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi:10.1038/nrg3833

Tam, L., McGlynn, L. M., Traynor, P., Mukherjee, R., Bartlett, J. M. S., and Edwards, J. (2007). Expression Levels of the Jak/stat Pathway in the Transition from Hormone-Sensitive to Hormone-Refractory Prostate Cancer. *Br. J. Cancer* 97, 378–383. doi:10.1038/sj.bjc.6603871

Tapia, G., and Diaz-Padilla, I. (2013). Molecular Mechanisms of Platinum Resistance in Ovarian Cancer. *Ovarian Cancer-A Clin. translational Update*, 205–223doi:10.5772/55562

Tian, D., Gu, Q., and Ma, J. (2016). Identifying Gene Regulatory Network Rewiring Using Latent Differential Graphical Models. *Nucleic Acids Res.* 44, e140. doi:10.1093/nar/gkw581

Uppal, K., Ma, C., Go, Y.-M., and Jones, D. P. (2018). Xmwas: a Data-Driven Integration and Differential Network Analysis Tool. *Bioinformatics* 34, 701–702. doi:10.1093/bioinformatics/btx656

van Zyl, B., Tang, D., and Bowden, N. A. (2018). Biomarkers of Platinum Resistance in Ovarian Cancer: what Can We Use to Improve Treatment. *Endocrine-related cancer* 25, R303–R318. doi:10.1530/erc-17-0336

Wang, C., Peng, G., Huang, H., Liu, F., Kong, D.-P., Dong, K.-Q., et al. (2018). Blocking the Feedback Loop between Neuroendocrine Differentiation and Macrophages Improves the Therapeutic Effects of Enzalutamide (Mdv3100) on Prostate Cancer. *Clin. Cancer Res.* 24, 708–723. doi:10.1158/1078-0432.ccr-17-2446

Wang, H., Fazayeli, F., Chatterjee, S., and Banerjee, A. (2014). Gaussian Copula Precision Estimation with Missing Values. *Artif. Intelligence Stat.*, 33, 978–986.

Wang, Y.-Y., Chen, W.-H., Xiao, P.-P., Xie, W.-B., Luo, Q., Bork, P., et al. (2017). Gear: A Database of Genomic Elements Associated with Drug Resistance. *Sci. Rep.* 7, 44085. doi:10.1038/srep44085

Wu, H.-C., Chang, C.-H., Wan, L., Wu, C.-I., Tsai, F.-J., and Chen, W.-C. (2006). Il-2 Gene C/t Polymorphism Is Associated with Prostate Cancer. *J. Clin. Lab. Anal.* 20, 245–249. doi:10.1002/jcla.20149

Wu, N., Huang, J., Zhang, X.-F., Ou-Yang, L., He, S., Zhu, Z., et al. (2019). Weighted Fused Pathway Graphical Lasso for Joint Estimation of Multiple Gene Networks. *Front. Genet.* 10, 623. doi:10.3389/fgene.2019.00623

Xu, T., Ou-Yang, L., Hu, X., and Zhang, X.-F. (2018). Identifying Gene Network Rewiring by Integrating Gene Expression and Gene Network Data. *Ieee/acm Trans. Comput. Biol. Bioinf.* 15, 2079–2085. doi:10.1109/tcbb.2018.2809603

Yuan, H., Xi, R., Chen, C., and Deng, M. (2017). Differential Network Analysis via Lasso Penalized D-Trace Loss. *Biometrika* 104, 755–770. doi:10.1093/biomet/asx049

Yuan, P., Miao, X. P., Zhang, X. M., Wang, Z. H., Tan, W., Zhang, X. R., et al. (2006). Association of the Responsiveness of Advanced Non-small Cell Lung Cancer to Platinum-Based Chemotherapy with P53 and P73 Polymorphisms. *Zhonghua Zhong Liu Za Zhi* 28, 107–110.

Yuan, Z., Cao, K., Lin, C., Li, L., Liu, H.-y., Zhao, X.-y., et al. (2011). The P53 Upregulated Modulator of Apoptosis (PUMA) Chemosensitizes Intrinsically Resistant Ovarian Cancer Cells to Cisplatin by Lowering the Threshold Set by Bcl-xL and Mcl-1. *Mol. Med.* 17, 1262–1274. doi:10.2119/molmed.2011.00176

Zhang, P., Liu, S. S., and Ngan, H. Y. S. (2012). Tap73-mediated the Activation of C-Jun N-Terminal Kinase Enhances Cellular Chemosensitivity to Cisplatin in Ovarian Cancer Cells. *PLoS ONE* 7, e42985. doi:10.1371/journal.pone.0042985

Zhang, X.-F., Ou-Yang, L., and Yan, H. (2017). Incorporating Prior Information into Differential Network Analysis Using Non-paranormal Graphical Models. *Bioinformatics* 33, 2436–2445. doi:10.1093/bioinformatics/btx208

Zhang, X.-F., Ou-Yang, L., Yan, T., Hu, X. T., and Yan, H. (2021). A Joint Graphical Model for Inferring Gene Networks across Multiple Subpopulations and Data Types. *IEEE Trans. Cybern.* 51, 1043–1055. doi:10.1109/TCYB.2019.2952711

Zhang, X.-F., Ou-Yang, L., Zhao, X.-M., and Yan, H. (2016). Differential Network Analysis from Cross-Platform Gene Expression Data. *Sci. Rep.* 6, 34112. doi:10.1038/srep34112

Zou, H., and Li, R. (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Ann. Stat.* 36, 1509–1533. doi:10.1214/009053607000000802

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 November 2021
doi: 10.3389/fgene.2021.751158

Identification of Intercellular Signaling Changes Across Conditions and Their Influence on Intracellular Signaling Response From Multiple Single-Cell Datasets

Mengqian Hao[1,2], *Xiufen Zou*[1,2] *and Suoqin Jin*[1,2]*

[1]School of Mathematics and Statistics, Wuhan University, Wuhan, China, [2]Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan, China

Identification of intercellular signaling changes across multiple single-cell RNA-sequencing (scRNA-seq) datasets as well as how intercellular communications affect intracellular transcription factors (TFs) to regulate target genes is crucial in understanding how distinct cell states respond to evolution, perturbations, and diseases. Here, we first generalized our previously developed tool CellChat, enabling flexible comparison analysis of cell–cell communication networks across any number of scRNA-seq datasets from interrelated biological conditions. This greatly facilitates the ready detection of signaling changes of cell–cell communication in response to any biological perturbations. We then investigated how intercellular communications affect intracellular signaling response by inferring a multiscale signaling network which bridges the intercellular communications at the population level and the cell state–specific intracellular signaling network at the molecular level. The latter is constructed by integrating receptor-TF interactions collected from public databases and TF-target gene regulations inferred from a network-regularized regression model. By applying our approaches to three scRNA-seq datasets from skin development, spinal cord injury, and COVID-19, we demonstrated the capability of our approaches in identifying the predominant signaling changes across conditions and the critical signaling mechanisms regulating target gene expression. Together, our work will facilitate the identification of both intercellular and intracellular dysregulated signaling mechanisms responsible for biological perturbations in diverse tissues.

Keywords: scRNA-seq data, intercellular communication, intracellular signaling, multiscale signaling network, dysregulated signaling, comparison analysis

OPEN ACCESS

Edited by:
Jiajun Zhang,
Sun Yat-sen University, China

Reviewed by:
Lin Wan,
Academy of Mathematics and Systems Science (CAS), China
Wei Vivian Li,
Rutgers, The State University of New Jersey, United States

***Correspondence:**
Suoqin Jin
sqjin@whu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *31 July 2021*
Accepted: *11 October 2021*
Published: *11 November 2021*

Citation:
Hao M, Zou X and Jin S (2021) Identification of Intercellular Signaling Changes Across Conditions and Their Influence on Intracellular Signaling Response From Multiple Single-Cell Datasets. Front. Genet. 12:751158. doi: 10.3389/fgene.2021.751158

INTRODUCTION

Cell–cell communication means that one cell sends a message to another cell through a medium to initiate cellular response of the target cell. The communication between cells plays a vital role in the development, physicology, and pathology of muticellular organisms. In this process, cells can communicate with and respond to neighboring or distant cells through ligand-receptor interactions by utilizing biochemical molecules, such as cytokines and growth factors.

Single-cell RNA-sequencing (scRNA-seq), which measures expression levels of a large number of genes across many cell types at a single-cell resolution, provides a great opportunity to study the cell–cell communication between interacting cells and the signaling response governed by intracellular gene regulatory networks (GRNs) (Almet, et al., 2021; Shao, et al., 2020). Moreover, identification of signaling changes across conditions is important for understanding how distinct cell states respond to evolution, perturbations, and diseases (Armingol, et al., 2021).

Although a number of computational methods have been recently developed to infer cell–cell communication by integrating scRNA-seq data with a prior ligand–receptor interaction database, most of these methods only focus on the intercellular communications in one biological condition (Almet, et al., 2021; Armingol, et al., 2021), lacking the capability of identifying signaling changes across conditions. We have recently developed a computational tool CellChat (Jin, et al., 2021) to identify dysregulated interactions by comparing cell–cell communication networks across conditions. However, CellChat focuses primarily on the comparison analysis between two datasets from two interrelated biological conditions. Other methods, including iTalk (Wang, et al., 2019) and Connectome (Raredon, et al., 2021), have also been developed recently to perform comparison analysis. With the increasing number of scNRA-seq datasets collected from multiple conditions, time points, and disease states, easy-to-use tools that can seamlessly identify signaling changes across any biological conditions from multiple scRNA-seq datasets are highly needed.

Understanding how cell–cell communication affects the gene expression of target cells via transcription factors (TFs) is crucial to understand how target cells respond to extracellular signals and eventually the functional role of cell–cell communication. However, there are only rudimentary efforts to link cell–cell communication to downstream response *via* GRNs (Browaeys, et al., 2020; Cheng, et al., 2021; Hu, et al., 2020; Sha, et al., 2020), such as NicheNet, scMLnet, and CytoTalk. NicheNet and scMLnet build GRNs by directly curating the interactions among ligands, receptors, TFs, and target genes from public databases, while CytoTalk infers GRN by calculating the mutual information between all pairs of genes without discriminating TFs from target genes. Constructing a multiscale signaling network, which links data-driven intercellular communications with intracellular TF-target regulations, still remains challenging, preventing the better understanding of cell type–specific response to cell–cell communication.

To address these limitations, we first generalized our previously developed R package CellChat to enable the comparison analysis of any number of datasets from multiple conditions, allowing ready identification of signaling changes across conditions. In addition, we infer a multiscale signaling network which integrates the ligand–receptor interactions inferred from CellChat, the receptor-TF interactions from public databases, and the TF- gene regulations from a mathematical optimization model taking into account the prior network information from public databases. Of note, we build cell type–specific networks from the integrated network by identifying enriched TFs and target genes based on the differential expression analysis. Therefore, our multiscale framework provides a clear understanding of how the upstream of the signaling pathway in cell–cell communication regulates the downstream target genes in a sequential way. We apply our approaches to three scRNA-seq datasets from mouse skin embryonic development, mouse spinal cord injury, and human COVID-19 infection. Applications not only demonstrate the capability of our methods but also provide novel insights into signaling mechanisms driving phenotype transitions.

RESULTS

Overview of Identifying Intercellular Signaling Changes Across Conditions and Their Link to Intracellular Signaling Response From Multiple scRNA-Seq Datasets

We first generalized our previously developed tool CellChat together with the R package, providing a more coherent and easy-to-use way to perform comparison analysis of cell–cell communication across conditions from any number of scRNA-seq datasets. Cellchat requires users to provide a scRNA-seq dataset (gene expression data across cells) with cell type labels as the input (**Figure 1A**). After receiving the input information, CellChat infers statistically and biologically significant cell–cell communication networks for each dataset. Compared to the original CellChat that was limited to the comparison analysis of only two datasets, the updated CellChat generalizes many existing functions, which enables systematical comparison analysis of intercellular communications across any number of scRNA-seq datasets. Of note, cell type compositions in different datasets do not need to be exactly the same. Moreover, by introducing a merged CellChat object from a list of CellChat objects, the updated CellChat allows the comparison analysis of cell–cell communication networks across all input datasets in a coherent and flexible fashion. Specifically, CellChat can identify the changes of the dominant sender and receiver in cell groups by comparing any two datasets using network centrality metrics such as out-degree and in-degree. CellChat can also identify the predominantly altered signaling pathways and ligand–receptor pairs by comparing the inferred communication probabilities and projecting the inferred cell–cell communication networks onto a shared low-dimensional space for any number of datasets. CellChat displays the results of comparative analysis of multiple datasets in a variety of intuitive visualization methods, such as scatter plots, heatmaps, bar plots, and bubble plots (**Figure 1A**).

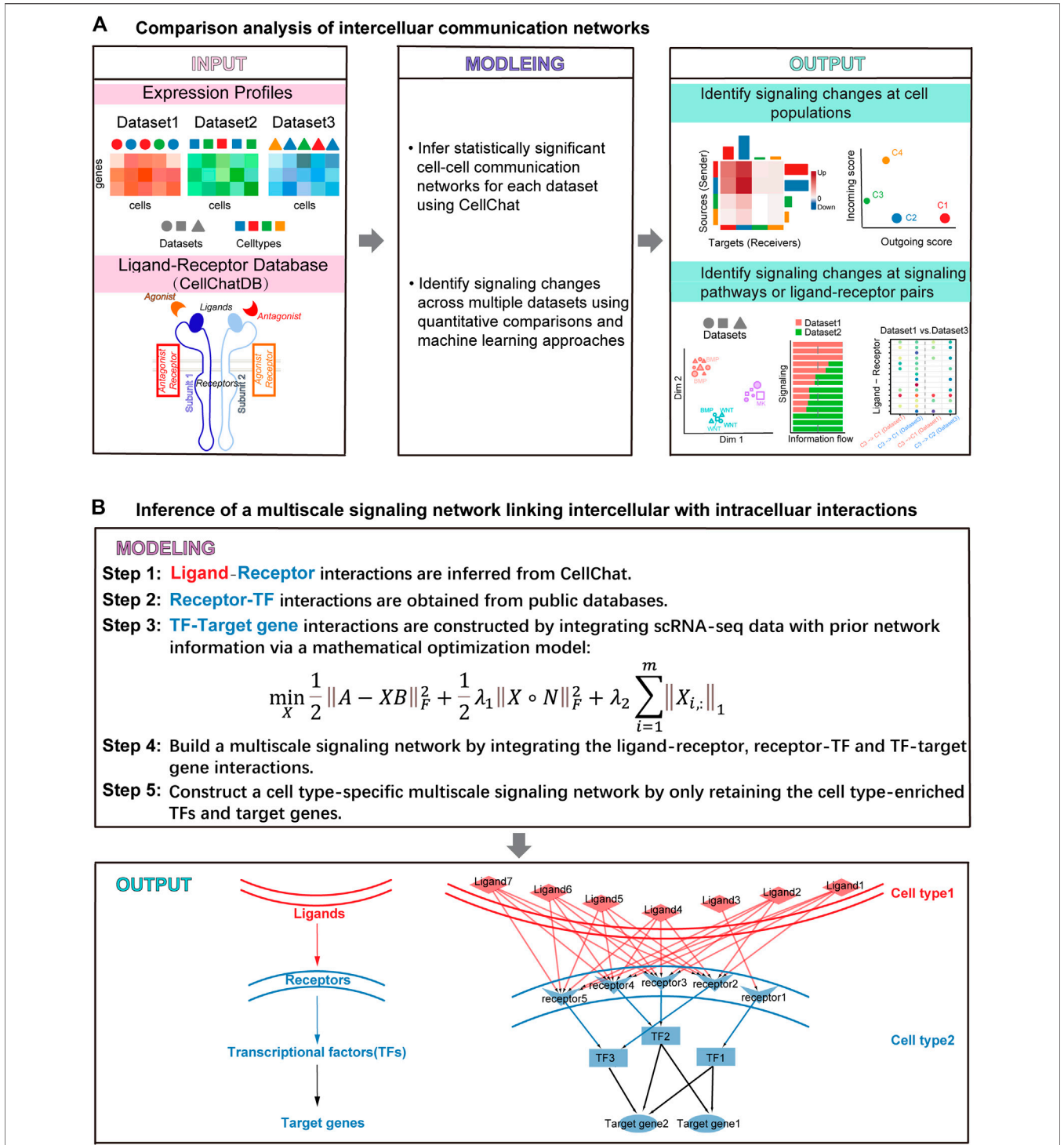


FIGURE 1 | Overview of identifying intercellular signaling changes across conditions and inferring the multiscale signaling network from multiple scRNA-seq datasets. **(A)** The generalized CellChat identifies intercellular signaling changes across conditions from multiple scRNA-seq datasets. CellChat requires the users to provide multiple datasets with cell type labels as input, where the cell types of different data sets may not be exactly the same. CellChat identifies biologically significant signaling pathways for each dataset separately and then performs comparative analysis across multiple datasets in a systematic and quantitative manner. CellChat identifies signaling changes across multiple datasets in terms of cell types and signaling pathways or ligand receptor pairs. Different plots are provided to allow ready comparison analysis. **(B)** Multiscale signaling network is inferred to link intercellular communication to intracellular signaling, which integrates the ligand–receptor interactions, receptor-TFs interactions, and TFs-target gene interactions.

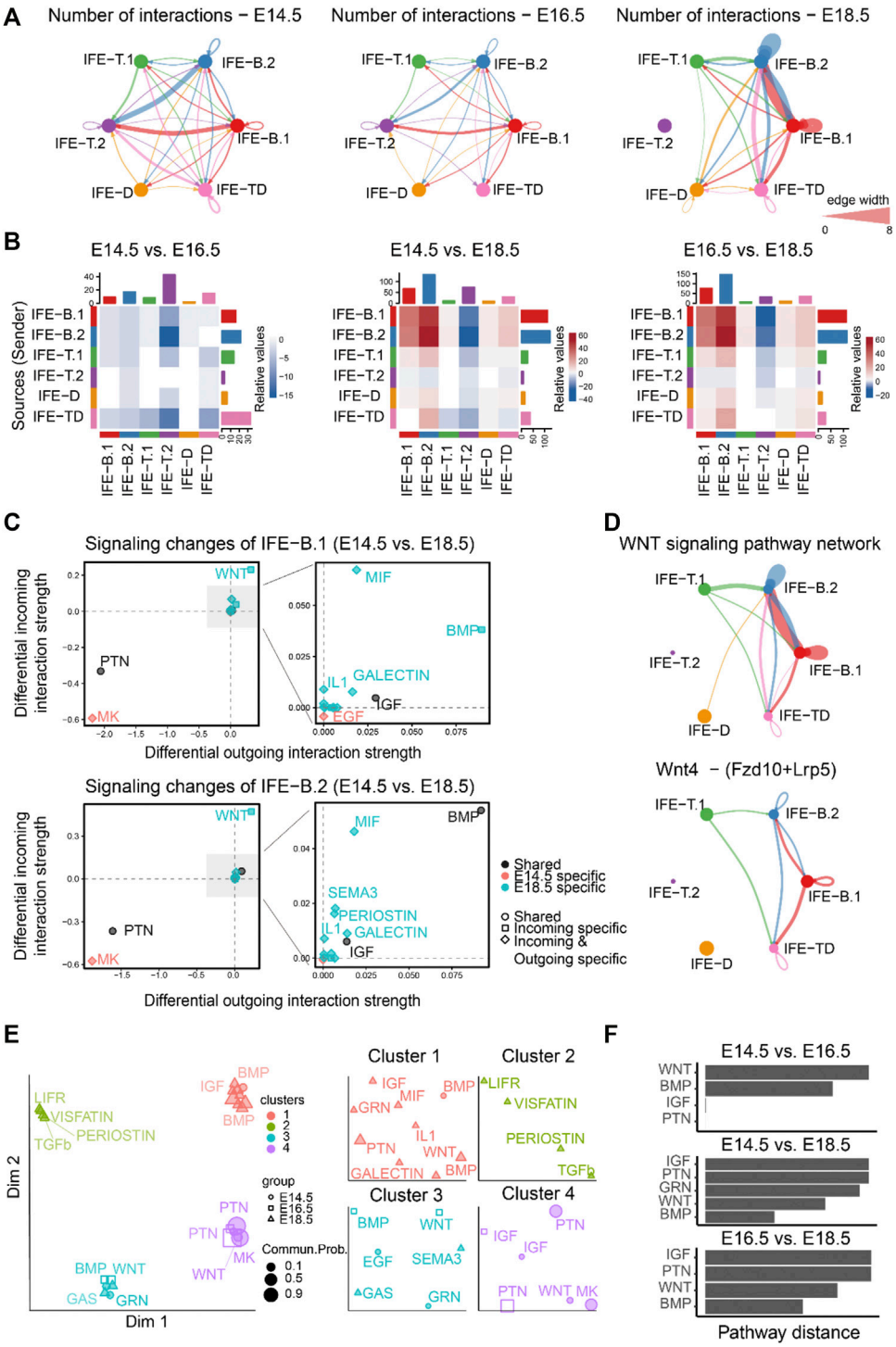


FIGURE 2 | Comparison analysis predicts WNT signaling as a predominant signaling change during mouse embryonic skin development. **(A)** The comparison of the total number of interactions among different cell populations between E14.5, E16.5, and E18.5. Edge width is proportional to the number of interactions, which assess how many ligand–receptor pairs contributing to the communication between two interacting cell populations. **(B)** Heatmap showing the differential number of interactions between E14.5, E16.5, and E18.5. In the color bar, red (or blue) represents increased (or decreased) signaling in the second dataset compared to the first one. **(C)** Identifying the specific signaling changes of IFE-B.1 and IFE-B.2 from E14.5 to E18.5. **(D)** The inferred WNT signaling pathway network and Wnt4 - (Fzd10 + Lrp5) signaling network at E18.5. **(E)** Projection and classification of signaling networks from E14.5, E16.5, and E18.5 onto a two-dimensional space based on the network similarity. Different shapes represent signaling networks from different developmental stages. **(F)** Computing the pathway distance of the signaling network from E14.5, E16.5, and E18.5 based on their Euclidean distance in the shared two-dimensional space.

We then build a multiscale signaling network by integrating the cell–cell communication between interacting cells (i.e., intercellular communication) with the downstream signaling inside target cells (i.e., intracellular signaling). The intercellular communication is given by CellChat while the intracellular signaling is inferred by constructing a gene–gene network linking receptors, TFs, and target genes. Specifically, the construction of a multiscale signaling network includes five steps. Step 1, the intercellular communication mediated by ligand–receptor interactions is obtained by CellChat. Step 2, the receptor-TF subnetwork is collected from a comprehensive database OmniPath (Türei, et al., 2016; Türei et al., 2021). Step 3, the TF-target gene subnetwork is inferred by integrating the TF activity data, the gene expression data, and the prior network information from the OmniPath database *via* a network-regularized regression model (MATERIALS AND METHODS). The TF activity is estimated based on their target gene expression in the scRNA-seq data using the widely used method DoRothEA (Garcia-Alonso, et al., 2019). Step 4, the multiscale signaling network is constructed by integrating the intercellular communication network, the receptor-TF network, and TF-target gene network, which links intercellular communications with intracellular signaling response. Step 5, the cell type–specific multiscale signaling network is finally constructed by only retaining the cell type–enriched TFs and target genes based on differential expression analysis (**Figure 1B**).

Together, our new approaches will advance our understanding of signaling mechanisms by identifying signaling changes that potentially drive phenotype transitions and by constructing multiscale signaling networks that imply how intercellular communications affect intracellular TFs to regulate target gene expression.

Comparison Analysis Predicts WNT Signaling as a Predominant Signaling Change During Mouse Embryonic Skin Development

To demonstrate the capability of our approaches in capturing predominant signaling changes across multiple time points, we first applied our generalized CellChat to our previously published mouse skin scRNA-seq datasets, which described epidermal development at three embryonic stages: E14.5, E16.5, and E18.5 (newborn) (Lin, et al., 2020). Unsupervised clustering identified five interfollicular epidermis (IFE) cell states: two basal cell states (IFE-B.1 and IFE-B.2), two transition cell states (IFE-T.1 and IFE-T.2), differentiated cells (IFE-D), and terminally differentiated cells (IFE-TD) (Lin, et al., 2020) (**Supplementary Figure S1A**).

To study how the cell–cell communication changes across different stages during mouse embryonic development, we first compared the number of inferred interactions among different cell populations among E14.5, E16.5, and E18.5 (**Figures 2A,B**). We observed slightly decreased cell–cell communication at E16.5 compared to E14.5, but significantly dynamic changes at E18.5 compared to both E14.5 and E16.5, suggesting dramatic signaling changes from E16.5 to E18.5 at the later embryonic stages. In particular, both outgoing and incoming signaling associated with IFE-B.1 and IFE-B.2 was predominantly increased at E18.5 compared to both E14.5 and E16.5. Surprisingly, our results showed that IFE-T.2 does not have any communication with any cell populations, which is likely due to the very few number of cells in IFET.2 at E18.5 (**Figure 2A** and **Supplementary Figure S1A**).

Moreover, to identify the signaling pathways contributing to the dramatic signaling changes of IFE-B.1 and IFE-B.2, we calculated the differential outgoing and incoming interaction strength of each signaling pathway between E14.5 and E18.5. For both IFE-B.1 and IFE-B.2, we observed WNT signaling as the most predominantly increased signaling at E18.5 compared to E14.5, as reflected by the largest differential outgoing and incoming interaction strength compared to other signaling pathways (**Figure 2C**), which was in agreement with the previous finding. In addition to WNT signaling, we also observed other increased signaling changes for both outgoing and incoming signaling including BMP, MIF, GALECTIN, and IL1, and decreased signaling including MK and PTN (**Figure 2C**). Attractively, our previous study experimentally showed that WNT-secreting stem cells play a central role in IFE self-renewal during homeostasis, which can inhibit the expansion of epidermal stem cells and the appearance of abnormal stem cell states (Lin, et al., 2020), in particular Wnt4 signaling. Indeed, we calculated the contribution of each ligand–receptor pair to the WNT signaling pathway and observed that Wnt4 - (Fzd10 + Lrp5) makes a relatively large contribution (**Supplementary Figure S1B**). By examining the gene expression levels of the ligand Wnt4 and its receptor Fzd10 and coreceptor Lrp5, IFE-B.1 and IFE-B.2 exhibited relatively high expression (**Supplementary Figure S1C**). Consistent with these observations, the inferred cell–cell communication networks of the WNT signaling pathway and the ligand–receptor pair Wnt4 - (Fzd10 + Lrp5) showed that IFE-B.1 and IFE-B.2 are the dominant signaling sources and targets at E18.5. In addition, IFE-T.1 and IFE-TD emerge as the signaling source and target, respectively, helping drive the complexity of WNT signaling.

We next investigated how the cell–cell communication architecture changes by projecting the inferred cell–cell communication networks from the three development stages onto a shared two-dimensional space based on whether they have similar signaling sources and targets (MATERIALS AND METHODS). This analysis classified all significant signaling pathways into four groups. Interestingly, the shared signaling pathways from two development stages were classified into different groups, such as WNT, BMP, and IGF (**Figures 2E,F**), suggesting that these pathways changed their cell–cell communication architecture during embryonic development.

Together, comparison analyses of the inferred cell–cell communication networks across the three embryonic

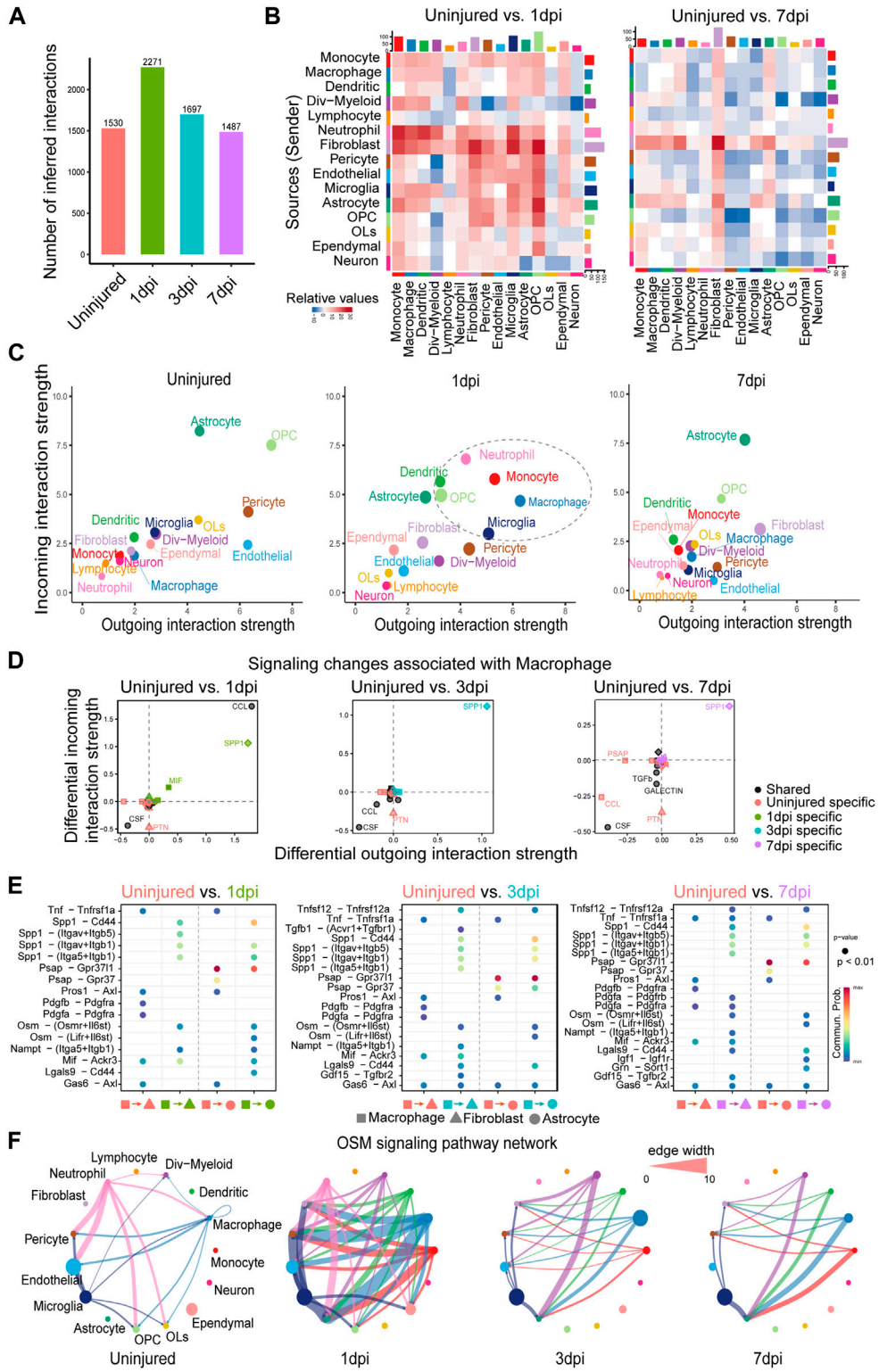


FIGURE 3 | Comparison analysis reveals myeloid cell–mediated signaling mechanisms and pinpoints the key time point of signaling changes in response to mouse spinal cord injury. **(A)** Comparison of the total number of interactions of the inferred cell–cell communication networks from uninjured, 1, 3, and 7dpi. **(B)** Heatmaps of the differential number of interactions between uninjured and 1dpi as well as uninjured and 7dpi, showing the outgoing and incoming signaling change of each cell group in a greater detail (The top colored bar plot represents the sum of each column of values displayed in the heatmap (incoming signaling). The right colored bar plot represents the sum of each row of values (outgoing signaling). **(C)** Scatter plots comparing the outgoing and incoming interaction strength in the 2D space among uninjured, 1dpi, and 7dpi. **(D)** Identifying signaling changes associated with macrophages by comparing uninjured with 1, 3, and 7dpi, respectively. **(E)** Identification of dysfunctional signaling by comparing the communication probabilities mediated by ligand–receptor pairs from macrophages to astrocytes and fibroblasts. **(F)** Circle plots displaying the inferred network of the OSM signaling pathway at uninjured, 1, 3, and 7dpi. Edge width is proportional to the inferred communication probabilities.

development stages suggest the dramatic signaling changes at the later stages and revealed WNT signaling as a predominant signaling change during mouse embryonic skin development.

Comparison Analysis Reveals Myeloid Cells-Mediated Signaling Mechanisms and Pinpoints the Key Time Point of Signaling Changes in Response to Mouse Spinal Cord Injury

Next, we demonstrate how our generalized CellChat can be applied in studying temporal changes of intercellular communications over four time points using a recently published mouse spinal cord injury sc-RNAseq dataset (Milich, et al., 2021). This dataset describes the wound healing process that occurs after spinal cord injury over four time points, including the uninjured and injured spinal cord at 1, 3, and 7 days postinjury (dpi). 66,176 cells were classified into 15 distinct cell groups: microglia, astrocytes, monocytes, macrophages, neutrophils, div-myeloid cells, dendritic cells, lymphocytes, oligodendrocytes (OLs), OPCs, neurons, fibroblasts, pericytes, ependymal cells, and endothelial cells.

We first compared the total number of interactions (i.e., the number of ligand–receptor pairs contributing to communication between any two interacting cell groups) that were inferred by CellChat over spinal cord injury. We found that the number of cell–cell communication was significantly increased at 1dpi after spinal cord injury, but afterwards decreased to its basal level by 7dpi (**Figure 3A**), suggesting that 1dpi was a critical time point where cell–cell communication between different cell types was significantly enhanced. To find out the interaction between which cell groups was significantly changed, we computed the differential number of interactions for both outgoing and incoming signaling of pairwise cell groups between any pair of two time points. We observed that the number of interactions between cell groups at 1dpi was mostly increased compared to the uninjured, while cell–cell communication at 3dpi and 7dpi exhibited a dynamic change with both increased and decreased interactions (**Figure 3B**). Interestingly, both outgoing signaling and incoming signaling of fibroblasts and astrocytes were consistently enhanced at 1dpi, 3dpi, and 7dpi compared to the uninjured, consistent with the known important role of fibroblasts in tissue repair (Plikus, et al., 2021) (**Figure 3B** and **Supplementary Figure S2A**).

In addition, we studied how the major signaling sources and targets changed after injury. Compared to the uninjured tissue, we found that both the outgoing and incoming interaction strength of several myeloid cell populations, including the macrophage, monocyte, neutrophil, microglia, and dendritic cells, were significantly increased at 1dpi, and the outgoing and incoming interaction strength of fibroblasts was increased at 3dpi and then, further enhanced at 7dpi (**Figure 3C** and **Supplementary Figure S2B**). These results agreed well with the previous findings: 1) At 1dpi, peripheral myeloid cells, mainly neutrophils and monocytes, migrate to the injury site and then enhance the innate immune response initiated by the microglia (Milich, et al., 2019); 2) Fibrosis is initiated at 3dpi and the number of fibroblasts reaches its peak at 7dpi (Zhu, et al., 2015). These two findings suggest the potential role of myeloid cells in initiating fibrosis after spinal cord injury.

To identify myeloid cell–mediated mechanisms of fibrosis, we examined signaling changes associated with macrophages by comparing its outgoing and incoming interaction strength of each signaling pathway at 1, 3, and 7dpi with the uninjured tissue. SPP1 signaling consistently exhibited the predominantly increased outgoing and incoming interaction strength at 1, 3, and 7dpi compared to the uninjured (**Figure 3D**), suggesting the important role of SPP1 signaling after spinal cord injury. This is consistent with the known neuroprotective roles of SPP1 and the worse histopathology and behavioral recovery in SPP1-knockout mice after spinal cord injury (Milich, et al., 2021). In addition, CCL signaling was also clearly increased at 1dpi compared to the uninjured (**Figure 3D**), which agreed with the innate immune response initiated by myeloid cells at 1dpi (Milich, et al., 2019). Furthermore, comparing the communication probabilities mediated by ligand–receptor pairs from macrophages to fibroblasts and astrocytes, we identified ligand-receptor pairs that were only enriched at 1, 3, and 7dpi, including SPP1 signaling such as Spp1 - (Itgav + Itgb5) and Spp1 - (Itga5+Itgb1) and OSM signaling such as Osm - (Osmr + Il6st) (**Figure 3E** and **Supplementary Figure S2C**). Consistent with our prediction, the previous study showed that OSM is a common mechanism by which fibroblasts and astrocytes are preferentially activated by monocyte/macrophage subtypes after spinal cord injury (Milich, et al., 2021). By examining the inferred cell–cell communication network at each time point, we found that OSM signaling was strongly activated with more signaling sources and a stronger interaction strength at 1dpi (**Figure 3F**). Compared to the uninjured tissue, other myeloid cells, including the monocyte, dendritic cells, and dividing myeloid cells (div-myeloid), emerged as new signaling sources, helping enhance the cell–cell communication driven by the macrophage. Notably, fibroblasts and astrocyte cells emerged as new signaling targets after injury, suggesting the myeloid cell–mediated signaling mechanisms of fibrosis. Taken together, our systematical comparison analysis pinpoints 1dpi as the key time point of signaling changes in response to spinal cord injury and reveals myeloid cell–mediated signaling mechanisms of fibrosis after mouse spinal cord injury.

Comparison Analysis Identifies Crucial Signaling Changes Responsible for Disease Severity Related to COVID-19

Due to the ongoing pandemic caused by the new coronavirus (SARS-CoV-2), it is of great significance to investigate the level of cell-to-cell communication in patients with different severity of diseases related to COVID-19. We used scRNA-seq data from 19 patients with COVID-19 and five SARS-CoV-2-

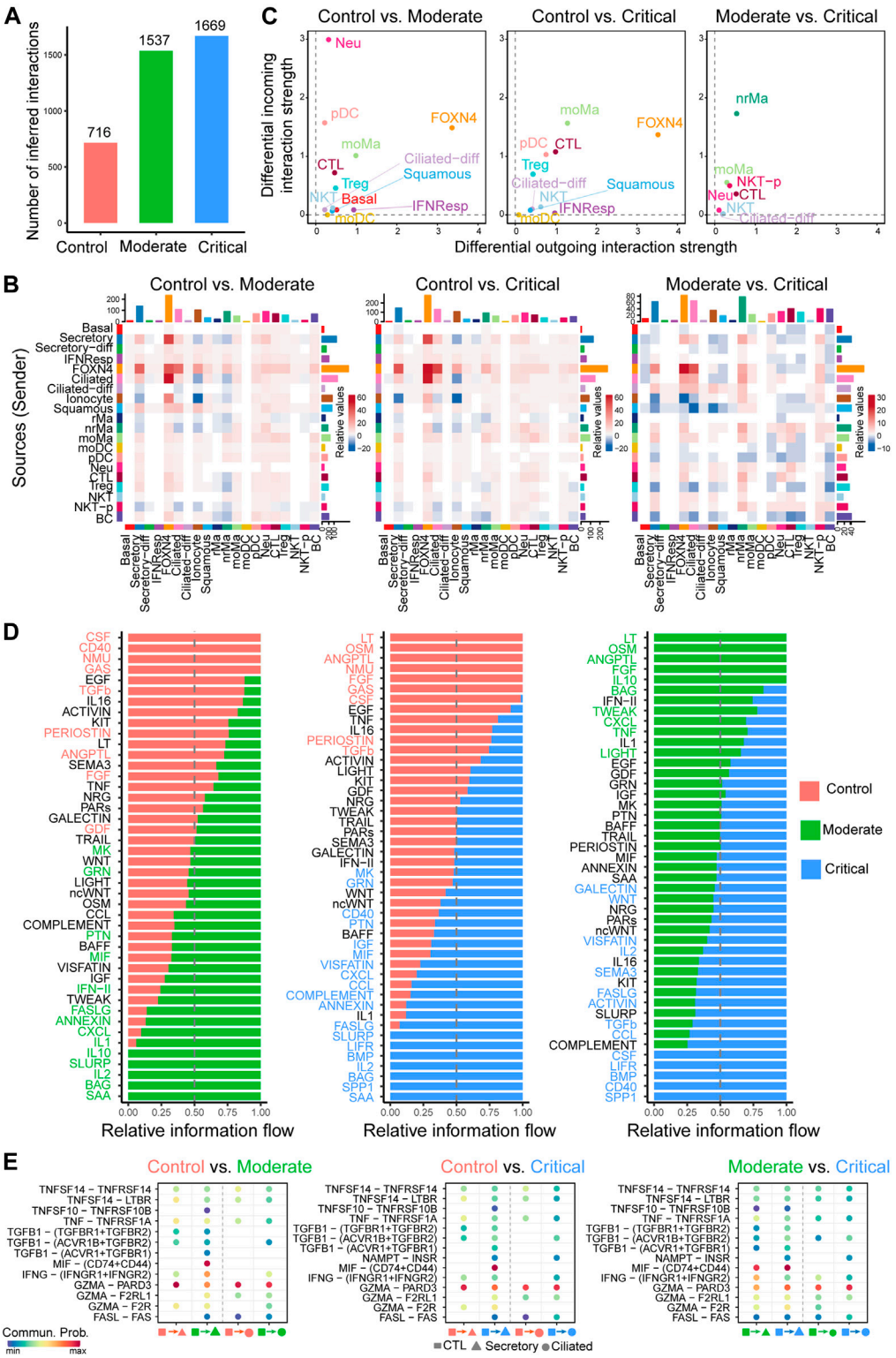


FIGURE 4 | Comparison analysis of cell–cell communication identifies major signaling changes in patients with COVID-19 across control, moderate, and critical cases. **(A)** Comparison of the number of interactions among different cell populations. **(B)** Differential number of interactions in the cell–cell communication network between control, moderate, and critical in greater details. **(C)** Signaling changes of the major cell groups that send or receive signals. Positive values in the differential outgoing (or incoming) interaction strength suggest the increased likelihood being sender (or receiver) in the second dataset compared to the first dataset. **(D)** The comparison of the signaling pathway based on the relative information flow between pairwise datasets. **(E)** Identifying altered ligand–receptor pairs from CTL to secretory and ciliated cells by comparing their communication probabilities between control, moderate, and critical.

negative donors with no signs of the disease. This dataset includes five control cases, eight moderate cases, and eleven critical cases. In the control, moderate, and critical samples, each contains 2,982, 82,814, and 49,804 cells. We performed downsampling of the moderate and critical samples by randomly taking 20,000 cells from each sample without losing any cell population. This dataset comprises 20 cell populations, including ciliated-diff cells (differentiating ciliated), secretory-diff cells (differentiating secretory), ciliated cells, FOXN4+ cells, squamous cells, secretory cells, cytotoxic T lymphocytes (CTL), natural killer T cells (NKT), B lymphocytes (BC), plasmacytoid dendritic cells (pDC), monocyte-derived macrophages (moMa), basal cells, proliferating NKT cells (NKT-p), IFNG-responsive cells (IFNRep), regulatory T cell (Treg), neutrophils (Neu), monocyte-derived dendritic cells (moDC), nonresident macrophages (nrMa), resident macrophages (rMa), and ionocytes.

After applying our generalized CellChat to the control, moderate, and critical samples separately, we calculated the total numbers of inferred interactions and observed an increased trend as the severity of the disease increases, with the highest interaction number of interactions detected in critical samples (**Figure 4A**). In more detail, we computed the differential number of interactions for both outgoing and incoming signaling of pairwise cell groups between different severities. Overall, the number of interactions was largely increased in moderate and critical samples compared to control, but exhibited dynamic changes when comparing moderate and critical (**Figure 4B**). Compared to control, the number of outgoing and incoming interactions of FOXN4, ciliated, and secretory cells in moderate and critical cases is higher. Compared to moderate cases, the outgoing signaling of FOXN4, ciliated, and some immune cells such as nrMa, moMa, Neu, CTL, NKT, and NKT-P was predominantly increased in critical cases (**Figure 4B**). Next, we examined the major source and target changes in different stages of COVID-19 by computing the differential outgoing and incoming differential interaction strength associated with each cell type (**Figure 4C**). Interestingly, compared to control, all cell types exhibited increased signaling in either outgoing or incoming signaling. In particular, FOXN4, CTL, moMa, pDC, and Treg in moderate and critical predominantly increased their outgoing and incoming interaction strength. CTL-, nrMa-, moMa-, Neu-, and NKT-associated signaling were further enhanced in critical compared to moderate.

We further focused on the specific signaling changes of two epithelial cell types: secretory and ciliated. Compared to control, certain chemokine and cytokine signaling pathways in moderate and critical were increased in their interaction strength (**Supplementary Figure S3A**). For the secretory-related signaling, CXCL, IFN-II, and IL1 increased either outgoing or incoming signaling; for the ciliated-related signaling, CCL, IFN-II, and IL2 increased either outgoing or incoming signaling. In addition, we compared the information flow (i.e., the sum of communication probabilities among all pairs of cell populations in the inferred network) for each signaling pathway between control, moderate, and critical samples (**Figure 4D**). We found that, compared to control, about half of the signaling pathways were highly enriched in moderate and critical (green and blue colors in left and middle panels in **Figure 4D**). These included many inflammatory signaling pathways such as IFN- II, CCL, CXCL, IL1, and IL2, suggesting that moderate and critical COVID-19 strongly trigger a series of inflammatory responses. Interestingly, compared to moderate, certain inflammatory response–related signaling were diminished in critical, such as OSM, IL10, TWEAK, CXCL, and LIGHT, while other inflammatory response–related signaling were enhanced in critical, such as IL2, IL16, CCL, LIFR, and CD40, suggesting that different inflammatory signaling likely play distinct roles in moderate vs. critical COVID-19.

Given the predominant signaling change of the immune cell CTL and epithelial cell secretory and ciliated, we investigate important ligand–receptor pairs sending from CTL cells to secretory and ciliated cells in moderate and critical. Compared to control, we observed that IFNG-(IFNGR1+IFNGR2) signaling was increased in both moderate and critical and TGFb-related signaling such as TGFB1-(ACVR1B + TGFBR2) was increased in critical compared to moderate (**Figure 4E**), suggesting the important role of IFN-II signaling in the interplay between immune cells and epithelial cells. Taken together, our comparison analysis revealed crucial signaling changes related to immune and epithelial cells and highlighted the ligand IFNG and its receptors IFNGR1 and IFNGR2 as critical enhanced signaling from CTL to secretory and ciliated cells, which might be responsible for disease severity related to COVID-19.

Multiscale Signaling Network Elaborates the Signaling Mechanisms of How SARS-CoV-2 Receptor ACE2 is Activated in Epithelial Lung Cells of Severe COVID-19

The binding of virus to the host receptor ACE2 greatly facilitates the infection of the mucosa of the upper respiratory by SARS-CoV-2. Therefore, the understanding of how ACE2 is activated in epithelial lung cells in patients with COVID-19 is crucial for therapeutic intervention of viral infection. To understand the role of cell–cell communication in activating ACE2 expression in the target cell, we constructed a multiscale signaling network by integrating the intercellular communications with the intracellular downstream signaling response (MATERIALS AND METHODS).

Our comparison analysis of cell–cell communication among control, moderate, and critical pinpoints the strong activation of cell–cell communication from the immune CTL cells to the epithelial secretory and ciliated cells mediated by IFN-II signaling in the moderate and critical compared to control (**Figure 4E**). By examining the inferred cell–cell communication network of the IFN-II signaling pathway (**Figure 5A**), we found that, compared to control, IFN-II

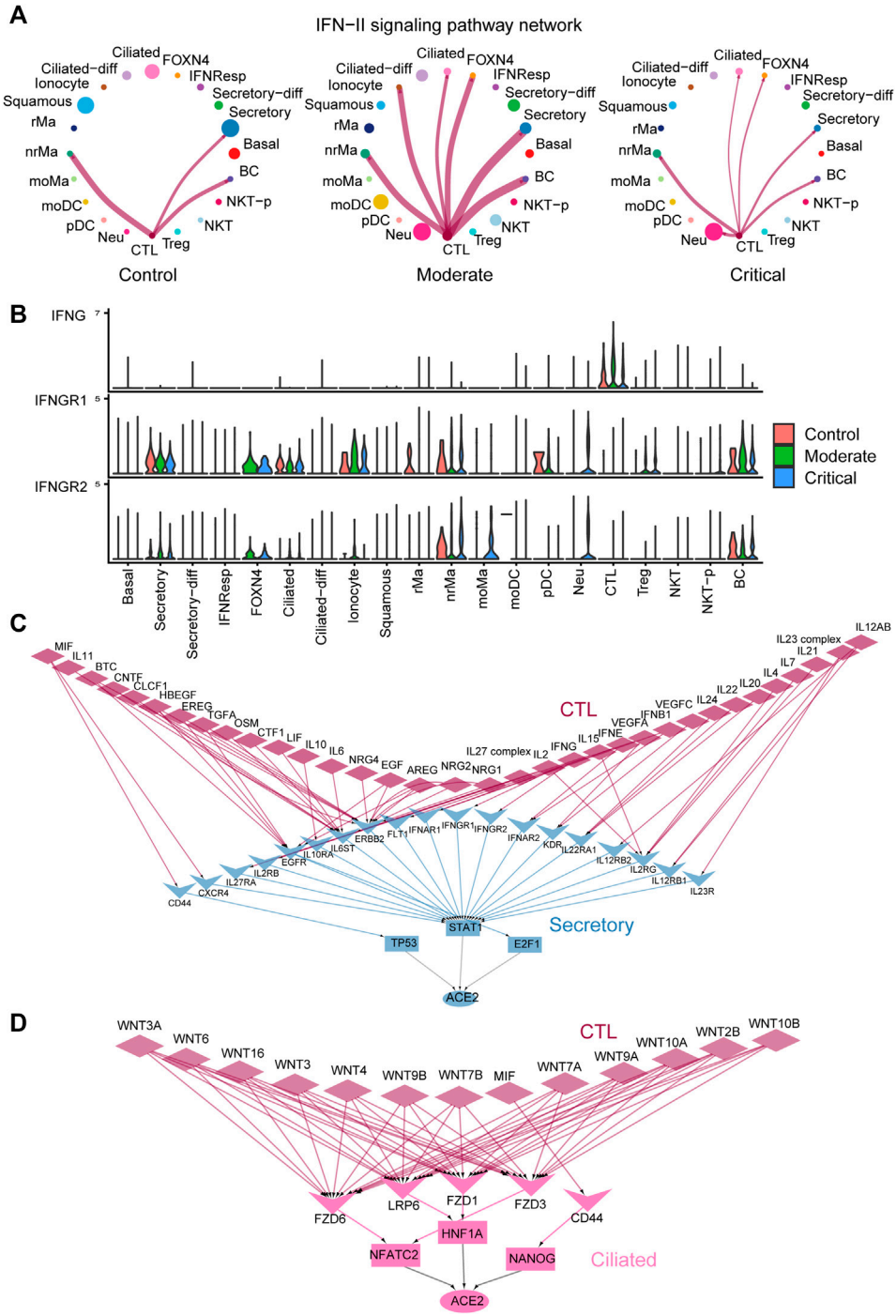


FIGURE 5 | Multiscale signaling network of CTL-to- secretory and CTL-to-ciliated reveals how intercellular communication activates the ACE2 expression *via* TFs in COVID-19. **(A)** Circle plots depicting the inferred IFN-II cell–cell communication networks between different cell groups in control, moderate, and critical. **(B)** Expression of IFN-II signaling–related genes such as IFNG, IFNGR1, and IFNGR2 in control, moderate, and critical COVID-19. **(C)** The inferred multiscale signaling network of CTL-to- secretory in critical COVID-19. **(D)** The inferred multiscale signaling network of CTL-to- ciliated in critical COVID-19.

signaling is strongly activated in moderate with a stronger interaction strength and more signaling targets. CTL is the dominant signaling source and FOXN4, ciliated, and ionocyte cells emerge as new signaling targets in moderate and critical (**Figure 5A**). Interestingly, the cell–cell communication strength is slightly diminished in critical compared to

moderate, possibly due to the relatively lower expression of the IFNG's receptors IFNGR1 and IFNGR2 in critical compared to moderate (**Figure 5B**).

Furthermore, we applied our computational framework of multiscale signaling network construction to study how CTL activates the ACE2 expression in the secretory and ciliated cells through the cell–cell communication. Therefore, we integrated the cell–cell communication network of CTL-to-secretory and CTL-to-ciliated with the downstream signaling network in secretory and ciliated cells, respectively. The downstream signaling network was constructed by integrating the receptor-TFs and TFs-target gene interactions (MATERIALS AND METHODS). Finally, we constructed two multiscale signaling networks for CTL-to-secretory and CTL-to-ciliated, respectively (**Figures 5C,D**). For the inferred multiscale signaling network of CTL-to-secretory, we observed many interferon, cytokines, and growth factors–related upstream ligands, such as IFNG, IFNB1, IFNE, IL6, IL10, IL2, IL17, OSM, IL4, IL7, VEGFA, EREG, TGFA, and EGF, as well as their corresponding receptors activated in secretory cells, such as IFNGR1, IFNGR2, IL6ST, IL2RG, KDR, and EGFR. Interestingly, our results showed that these ligand–receptor pairs activated three TFs, including STAT1 as the major activator and E2F1 and TP53 as the minor activators, and ACE2 can be activated by these three TFs. These results suggested that STAT1 was the major regulator to activate ACE2 in secretory cells, which is consistent with the previous finding (Chua, et al., 2020) and the known important role of the JAK-STAT signaling pathway during viral infection. Surprisingly, the inferred multiscale signaling network of CTL-to-ciliated showed that WNT signaling was highly activated in ciliated cells, which triggers the activation of three TFs including NFATC2, HNF1A, and NANOG and further activates the downstream target gene ACE2 expression (**Figure 5D**). Interestingly, previous studies showed that HNF1A is a master regulator of ACE2, and overexpression of HNF1A and ACE2 indicates greater risk of death or cardiovascular disease events (Narula, et al., 2020). In addition, NFATC2 is the predominant NFAT family members in the peripheral immune system and may be as a potential marker related to lung damage (Maremanda, et al., 2020). These results suggest the potential role of these TFs in regulating ACE2 expression in ciliated cells and might be considered as new therapeutic targets. Taken together, our multiscale signaling framework helps to elaborate the signaling mechanisms of how the SARS-CoV-2 receptor ACE2 is activated by TFs in epithelial lung cells of severe COVID-19.

Comparison CellChat With Other Cell-Cell Communication Tools

The characteristics of CellChat and its comparison with other tools, including iTalk, Connecctome, and NicheNet, are summarized in **Supplementary Figure S4A**. Briefly, compared to these three tools, the updated CellChat is the only easy-to-use tool that can seamlessly identify signaling changes across any number of scRNA-seq datasets. NicheNet does not perform comparison analysis across distinct datasets. These three tools do not consider the multisubunit structure of ligand–receptor complexes and membrane-bound stimulatory and inhibitory cofactors, which are necessary for certain ligand–receptor binding. Moreover, iTalk and Connectome do not infer the intracellular signaling network.

Since our previous study has already performed comparison analysis with iTalk (Jin et al., 2021) and NicheNet does not explicitly infer the cell–cell communication network, here we only compare CellChat with Connectome in their ability of identifying signaling changes across conditions. We aimed to identify signaling changes responsible for disease severity related to COVID-19. We found that CellChat produced upregulated and downregulated signaling genes that were more differentially expressed compared to Connectome, as reflected by a higher avg [log2(FC)] and –log10 (p_val_adj) of genes in the predicted ligand-receptor pairs (**Supplementary Figure S4B**). This result suggests that CellChat inferred more significant ligand–receptor interactions that were changed across conditions. By examining the list of inferred signaling pathways, interestingly, Connectome did not produce the IFN-II signaling while CellChat did. This signaling pathway has been shown to be strongly activated in moderate and critical compared to control during COVID infection (Chua et al. 2020). This result indicates CellChat's ability in predicting dysfunctional signaling pathways across conditions.

MATERIALS AND METHODS

CellChat requires gene expression data of cells as the user inputs and models the probability of cell–cell communication by integrating gene expression with prior knowledge of the interactions between signaling ligands, receptors, and their cofactors. Upon inferring the intercellular communication network, CellChat provides functionality for further data exploration, analysis, and visualization (Jin, et al., 2021). Compared to the original CellChat, here we made two important additions. First, the updated CellChat enables systematical comparison analysis of intercellular communication between interacting cells across any number of scRNA-seq datasets rather than limiting to two datasets. In this way, significant signaling changes across multiple conditions or time points can be presented in an intuitive way. Second, the updated CellChat is able to infer the multiscale signaling network linking intercellular communication with intracellular downstream signaling, which helps to better understand how the upstream of the signaling pathway in intercellular communication affects intracellular TFs to regulate the target gene expression.

Comparison Analysis of Intercellular Communication Between Interacting Cells Across Multiple Datasets

We generalized some functions and analysis in our previously developed R package CellChat, which can then be used for comparative analysis across multiple datasets. Here, we briefly described several key functionalities in the updated CellChat R package.

Identification of Important Signaling Sources and Targets in the Intercellular Communication Networks

We identified the dominant signaling sources and targets by defining the outgoing and incoming interaction strength as the out-degree and in-degree centrality metrics in the weighted cellular communication network, where the edge weights are assigned by the communication probabilities computed from CellChat (Jin, et al., 2021). The in-degree refers to the sum of the communication probabilities of incoming signaling to a cell group, while the out-degree is computed as the sum of communication probabilities of the outgoing signaling from a cell group. In this way, we can study the detailed changes in the outgoing and incoming signaling across all significant pathways.

Identification of Altered Signaling Pathways by Comparing the Information Flow of Each Signaling Pathway

The information flow for each signaling pathway is defined by the sum of communication probabilities among all pairs of cell groups in the inferred network (that is, the total weights in the network). We can compare the total information flow in the cell–cell communication network of each signaling pathway across different datasets under different conditions, leading to the identification of changes in important signaling pathways.

Identification of Signaling Networks With Architecture Difference Across Multiple Datasets Based on Their Network Similarity

CellChat quantifies the similarity of multiple cellular communication networks using structural similarity and functional similarity and performs joint manifold learning and classification of the inferred communication networks based on the computed similarity to identify signaling networks with a certain difference. Here, we focus on the functional similarity, which is calculated by using Jaccard similarity on the basis of the overlap of the major targets and sources in communications defined by:

$$S = \frac{E(G) \cap E(G')}{E(G) \cup E(G') - E(G) \cap E(G^{\wedge\prime})}, \quad (1)$$

where G and G' are two signaling networks, and $E(G)$ is the set of communications in signaling network G. The higher the functional similarity, the more similar the major senders and receivers are, which means that the two signaling pathways or two ligand–receptor pairs exhibit more similarity. Therefore, two cell-cell communication networks showing less functional similarity suggest that they change their signaling sources and targets across different datasets, implying the difference in network architecture.

Inference of Multiscale Signaling Network by Integrating Intercellular Communication with Intracellular Signaling Network

The construction of a multiscale signaling network includes the following five steps.

Step 1: Construction of the ligand–receptor subnetwork.

A very important way of information transmission between cells is the interaction between ligands and receptors on the cell surface. The ligand–receptor subnetwork is obtained by applying CellChat to the scRNA-seq data, which infers the biologically significant cell–cell communication network mediated by ligand–receptor interactions based on the database CellChatDB of ligand-receptor pairs in human and mice (Jin, et al., 2021).

Step 2: Construction of the receptor-TF subnetwork.

From the public databases, we get the receptor-TF prior network from the OmniPath database (Türei, et al., 2016; Türei et al., 2021), "kinaseextra" and "pathwayextra" using OmnipathR package (https://github.com/saezlab/OmnipathR).

Step 3: Construction of the TF-target gene subnetwork.

We focused on the cell type–specific signaling network and thus first identified enriched genes and TFs in each cell group. The nonparametric Wilcoxon rank sum test in Seurat v.3 (FindAllMarkers function) was used to perform differential gene expression analysis (min.pct = 0.25, logfc. threshold = 0.25). Genes were considered as enriched genes with an adjusted p-value < 0.05. To better model the relationship between TFs and their target genes, we estimated TF activity based on the target's mRNA expression level from scRNA-seq data using DoRothEA (Garcia-Alonso, et al., 2019) since TF activity is difficult to measure directly and it may be possible to infer changes in the TF activity level from changes in the expression levels of the TF's target genes. We then identified the enriched TFs in certain cell groups using the differential expression analysis based on the computed TF activity data.

To better infer the TF-target gene regulatory network, we integrated TF-target gene interactions from public databases with scRNA-seq data. We selected TF-target gene interactions with high confidence levels A, B, and C from the OmniPath database. Then, the inference of the TF-target gene regulatory network can be formulated as the following mathematical optimization problem

$$\min_X \frac{1}{2}\|A - XB\|_F^2 + \frac{1}{2}\lambda_1 \|X \circ N\|_F^2 + \lambda_2 \sum_{i=1}^{m} \|X_{i,:}\|_1, \quad (2)$$

where X is the TF-target regulatory network we need to infer. A is the target gene expression matrix (rows are target genes and columns are cells. B is the TF activity matrix (rows are TFs and columns are cells). N is the prior TF-target network from the public database (rows are target genes and columns are TFs). The value of each element N_{ij} is 0 or 1, where 0 means that there is no priori connecting edge between TF_j and

$Target\ gene_i$, and 1 indicates that there is a prior connecting edge. ∘ represents dot product. The last term constrains the sum of the absolute value of all link's weight coefficients, which can reduce the complexity of the model and make the network sparse, leading to more biologically explanatory results. Here, we choose the two regularization parameters λ_1 and λ_2 as 50 and 10, respectively.

We used the ADMM algorithm to efficiently solve this optimization problem. We rewrite the optimization problem as:

$$\min_{X,Z} \frac{1}{2} \| A - XB \|_F^2 + \frac{1}{2}\lambda_1 \| Z \circ N \|_F^2 + \lambda_2 \sum_{i=1}^{m} \| Z_{i,} \|_1. \quad (3)$$

Subject to $X - Z = 0$

The augmented Lagrangian with penalty parameter $t > 0$ is:

$$L_t(\mathrm{X}, \mathrm{Z}, \mathrm{Y}) = \frac{1}{2} \| A - XB \|_F^2 + \frac{1}{2}\lambda_1 \| Z \circ N \|_F^2 + \lambda_2 \sum_{i=1}^{m} \| Z_{i,} \|_1 + tY^T (X - Z) + \frac{t}{2}. \quad (4)$$

We then solved this optimization problem by following the update rules and stop criterion.

Update rules:

1) Update X:

$$X_k = arg \min_X L_t(\mathrm{X}, Z_{k-1}, \mathrm{Y}_{k-1}) \quad (5)$$

2) Update Z:

$$Z_k = arg \min_Z L_t(X_k, Z, \mathrm{Y}_{k-1}) \quad (6)$$

3) Update Y:

$$Y_k = Y_{k-1} + t(X_k - Z_k) \quad (7)$$

Stop criterion:

dual residual: $S_k = -t(Z_k - Z_{k-1})$

primal residual: $R_k = X_k - Z_k$

iteration stops when both $||R_k||_F$ and $||S_k||_F$ values become smaller than ϵ^{pri} and ϵ^{dual}, respectively,

$$||R_k||_F < \epsilon^{pri},$$
$$||S_k||_F < \epsilon^{dual},$$

where

$$\epsilon^{pri} = \sqrt{n}\epsilon^{abs} + \epsilon^{rel} \max\{||X_k||_F, ||-Z_k||_F\},$$
$$\epsilon^{dual} = \sqrt{m}\epsilon^{abs} + \epsilon^{rel} ||Y_k||_F.$$

After obtaining the solution X, we determined the weight of the network by considering another proportionality-based association measure "propr" (Quinn, et al., 2017), which was shown to perform very well in inferring gene networks across multiple scRNA-seq datasets and technologies (Skinnider, et al., 2019). We then defined the weights in the TF-target gene network as

$$X_{new} = \omega \cdot X_{model} + (1 - \omega) \cdot X_{propr}. \quad (8)$$

Here, we took the value of ω as 0.7. The weighted average is an ensemble strategy that has been widely used in many other studies. We also performed comparison analysis of networks inferred using weighted average X_{new} and using X_{model}. By using a prior network from public databases as a reference, we computed true positive rate (TPR), false positive rate (FPR), and the area under the ROC curve (AUC) and showed that the network inferred with the weighted average produces better results than X_{model} (**Supplementary Figure S5B**).

Step 4: Integration of intercellular communication network with intracellular signaling network.

We subset the receptor-TF network by only retaining receptors in the intercellular communication network and TFs in the TF-target gene network. Once we constructed the intercellular communication network mediated by ligand–receptor interactions, the receptor-TF network, and TF-target gene network, we integrated them together to obtain a multiscale signaling network, linking the intercellular communication network with intracellular signaling network.

Step 5: Inference of cell type–specific multiscale signaling network.

Finally, we build the cell type–specific multiscale signaling network based on whether the TFs and target genes were enriched in certain cell types based on the differential expression analysis. Of note, we construct the downstream intracellular signaling network for each dataset separately. To visualize the inferred network, we only retained the top 25 edges based on the inferred edge weights.

Robustness Analysis of Regularization Parameters

Our model is not sensitive to the regularization parameters within certain ranges. To demonstrate this point, we conducted robustness analysis and varied the regularization parameter values within a certain range to explore the robustness of our model. Specifically, we varied the regularization parameters λ_1 from 30 to 70 with an increment of 10 and λ_2 from 5 to 15 with an increment of 5, respectively. We then computed the residual value of the model using five-fold cross-validation under each parameter combination. We observed that the residual value exhibited a slight fluctuation (**Supplementary Figure S5A**), suggesting that our inference is relatively robust.

Single-Cell RNA-Seq Datasets, Data Preprocessing, and Analysis

Mouse Embryonic Skin scRNA-Seq Datasets

Interfollicular epidermis (IFE) covers the surface of the animal body and is a keratinized stratified squamous epithelium. The datasets (GEO accession codes: GSE154579) we used were published from our previous study (Lin, et al., 2020), containing three developmental stages: E14.5, E16.5, and E18.5 (newborn). The IFE cells were classified into six cell states: basal cells (IFE-B.1 and IFE-B.2), transition cells (IFE-T.1 and IFE-T.2), differentiated cells (IFE-D), and terminally differentiated cells (IFE-TD). Normalized data were used for all the analyses.

Mouse Spinal Cord Injury Datasets

Spinal cord injury is the most serious complication of spinal cord injury, often leading to severe dysfunction of the limbs below the injured segment and triggers multiple processes. The published spinal cord injury mouse datasets were downloaded from GEO (accession codes: GSE162610) and included a total of 66,176 cells from the uninjured and 1, 3, and 7dpi tissue (Zhu, et al., 2015). The original study classified these cells into 15 distinct cell groups: microglia, astrocytes, monocytes, macrophages, neutrophils, div–myeloid cells, dendritic cells, lymphocytes, oligodendrocytes, OPCs, neurons, fibroblasts, pericytes, ependymal cells, and endothelial cells. Normalized data were used for all the analyses.

COVID-19 Datasets

The processed transcriptomic data of 135,600 cells from patients and control patients with no signs of disease with COVID-19 were downloaded from FigShare: https://doi.org/10.6084/m9.figshare.12436517. This dataset includes eight moderate cases, eleven critical cases, and five control cases (According to the World Health Organization (WHO) guidelines, the severity of the disease is classified) (Chua, et al., 2020). In the control, moderate, and critical samples, each contains 2,982, 82,814, and 49,804 cells. We performed downsampling analysis on the moderate and critical cases with a maximum of 20,000 cells to reduce computational cost. This dataset contains 20 cell types, including ciliated-diff cells (differentiating ciliated), secretory-diff cells (differentiating secretory), ciliated cells, FOXN4+ cells, squamous cells, secretory cells, cytotoxic T lymphocytes (CTL), natural killer T cells (NKT), B lymphocytes (BC), plasmacytoid dendritic cells (pDC), monocyte-derived macrophages (moMa), basal cells, proliferating NKT cells (NKT-p), IFNG-responsive cells (IFNRep), regulatory T cell (Treg), neutrophils (Neu), monocyte-derived dendritic cells (moDC), nonresident macrophages (nrMa), resident macrophages (rMa), and ionocytes. To infer the intracellular signaling network in secretory cells, ciliated cells, and CTL, we only used the top 20 marker genes and the top 50 TFs associated with each cell population based on the differential expression analysis. Normalized data were used for all the analyses.

DISCUSSION

In this study, we generalized our previously developed tool CellChat to perform comparison analysis of cell–cell communication across multiple conditions or time points and established an optimization-based framework to construct a multiscale signaling network linking intercellular communication with intracellular downstream signaling response. This comparative analysis of the interactions between cell types across different biological conditions is essential for a biologically meaningful understanding of the role of cell–cell communication from scRNA-seq data. We demonstrated the effectiveness of our proposed approaches by studying the signaling changes across three mouse embryonic developmental stages, four time points after mouse spinal cord injury, and patients with different COVID-19 severities (i.e., control, moderate, and critical cases).

We found that our predictions can recapitulate known biology to a substantial degree. For example, the prediction of the WNT signaling pathway as the predominant signaling change during mouse embryonic development is in agreement with our previous finding that WNT signaling can inhibit the expansion of epidermal stem cells and the appearance of abnormal stem cell states during epidermal differentiation (Lin, et al., 2020). Our predictions also reveal many signaling changes that recapitulate previous findings or known biology during mouse spinal cord injury, such as the increased myeloid cell–associated interactions at 1dpi and enhanced OSM and SPP1 signaling, suggesting the important signaling mechanisms of fibrosis mediated by myeloid cells during wound healing after spinal cord injury. We found that the IFN-II signaling pathway has changed significantly in the patients of COVID-19 and can activate the master regulator STAT1 to regulate the downstream ACE2 expression in the secretory cells.

Although recent studies have developed different computational methods to investigate cell–cell communication, our study adds important understanding of the cell–cell communication in several aspects. On the one hand, we provide generalized functions in the CellChat R package for comparative analysis of any number of datasets and even for datasets with not exactly the same cell type compositions under different conditions. It compares the number of interactions; it also identifies changes of major sources and targets in cell groups and changes in signaling pathways and ligand–receptor pairs. The advantage is that the single-cell datasets used for comparative analysis can be any number, not just limited to the comparison between two datasets. Furthermore, we defined signaling similarity by computing the Jaccard similarity between the inferred cell–cell communication networks across different datasets. Our current strategy that combines clustering analysis can help to identify signaling networks that show a relatively large difference in network architecture if they are located in different clusters and far away from each other in the low-dimensional space. However, considering more advanced methods such as statistical tests could likely improve such analysis. We also identified significant changes in senders and receivers of each signaling pathway using network centrality measures such as out-degree and in-degree to characterize the outgoing and incoming interaction strength. Finally, we can use various forms of graphics as output to visualize our results, making the results more intuitive.

On the other hand, we proposed a mathematical optimization model that can infer the TF-target gene network by adding priori network information as a penalty term. Previous studies have also focused on the downstream signaling transduction of cell communication, but these methods like NicheNet and scMLnet primarily use prior network information from public databases, lacking the integration of single cell data in a coherent way. In contrary, our work lies in the integration of mathematical optimization models and prior network information based on a data-driven approach. Although previous studies showed that incorporating

such prior information as a network constraint can improve the model performance (e.g., Zhang and Zhang, 2020), reconstruction of the TF-target network directly from single-cell data using a more advanced method such as scLink (Li and Li, 2021) will be likely helpful to build a better multiscale signaling network. Furthermore, we extracted the cell type–specific network based on differential expression analysis and integrated with the upstream intercellular communication network to form a multiscale cellular communication network. In this way, the network we build will likely be more precise and more biologically explanatory.

As single-cell multi-omics data is becoming more common (Argelaguet, et al., 2021; Jin, et al., 2020; Zhang and Nie, 2021), the emergence of these data is a challenging opportunity to build a more systematic cellular communication network. In addition, spatial transcriptomics provide additional information on the cell location (Longo, et al., 2021). Integrating spatial location with scRNA-seq data will likely reduce the false positive inference of cell–cell communication.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

SJ designed research. MH and SJ performed research. MH, XZ, and SJ wrote the paper. SJ and XZ supervised research. All authors approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grant No 11801207 and 11831015, National Key Research and Development Program of China under Grant No 2018YFC1314600, and Natural Science Foundation of Hubei Province under Grant No 2019CFA007.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.751158/full#supplementary-material

REFERENCES

Almet, A. A., Cang, Z., Jin, S., and Nie, Q. (2021). The Landscape of Cell-Cell Communication Through Single-Cell Transcriptomics. *Curr. Opin. Syst. Biol.* 26, 12–23. doi:10.1016/j.coisb.2021.03.007

Argelaguet, R., Cuomo, A. S. E., Stegle, O., and Marioni, J. C. (2021). Computational Principles and Challenges in Single-Cell Data Integration. *Nat. Biotechnol.* 39, 1202–1215. doi:10.1038/s41587-021-00895-7

Armingol, E., Officer, A., Harismendy, O., and Lewis, N. E. (2021). Deciphering Cell-Cell Interactions and Communication From Gene Expression. *Nat. Rev. Genet.* 22 (2), 71–88. doi:10.1038/s41576-020-00292-x

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: Modeling Intercellular Communication by Linking Ligands to Target Genes. *Nat. Methods.* 17 (2), 159–162. doi:10.1038/s41592-019-0667-5

Cheng, J., Zhang, J., Wu, Z., and Sun, X. (2021). Corrigendum to: Inferring Microenvironmental Regulation of Gene Expression from Single-Cell RNA Sequencing Data Using scMLnet With an Application to COVID-19. *Brief Bioinform.* 22 (2), 1511–1512. doi:10.1093/bib/bbab015

Chua, R. L., Lukassen, S., Trump, S., Hennig, B. P., Wendisch, D., Pott, F., et al. (2020). COVID-19 Severity Correlates with Airway Epithelium-Immune Cell Interactions Identified by Single-Cell Analysis. *Nat. Biotechnol.* 38 (8), 970–979. doi:10.1038/s41587-020-0602-4

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and Integration of Resources for the Estimation of Human Transcription Factor Activities. *Genome Res.* 29 (8), 1363–1375. doi:10.1101/gr.240663.118

Hu, Y., Peng, T., Gao, L., and Tan, K. (2020). CytoTalk: De Novo Construction of Signal Transduction Networks Using Single-Cell RNA-Seq Data. *Sci. Adv.* 7 (16), eabf1356. doi:10.1101/2020.03.29.014464

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., et al. (2021). Inference and Analysis of Cell-Cell Communication Using CellChat. *Nat. Commun.* 12 (1), 1088. doi:10.1038/s41467-021-21246-9

Jin, S., Zhang, L., and Nie, Q. (2020). scAI: an Unsupervised Approach for the Integrative Analysis of Parallel Single-Cell Transcriptomic and Epigenomic Profiles. *Genome Biol.* 21 (1), 25. doi:10.1186/s13059-020-1932-8

Li, W. V., and Li, Y. (2021). scLink: Inferring Sparse Gene Co-expression Networks From Single-Cell Expression Data. *Genomics Proteomics Bioinformatics.* doi:10.1016/j.gpb.2020.11.006

Lin, Z., Jin, S., Chen, J., Li, Z., Lin, Z., Tang, L., et al. (2020). Murine Interfollicular Epidermal Differentiation Is Gradualistic With GRHL3 Controlling Progression From Stem to Transition Cell States. *Nat. Commun.* 11 (1), 5434. doi:10.1038/s41467-020-19234-6

Longo, S. K., Guo, M. G., Ji, A. L., and Khavari, P. A. (2021). Integrating Single-Cell and Spatial Transcriptomics to Elucidate Intercellular Tissue Dynamics. *Nat. Rev. Genet.* 22, 627–644. doi:10.1038/s41576-021-00370-8

Maremanda, K. P., Sundar, I. K., Li, D., and Rahman, I. (2020). Age-Dependent Assessment of Genes Involved in Cellular Senescence, Telomere, and Mitochondrial Pathways in Human Lung Tissue of Smokers, COPD, and IPF: Associations With SARS-CoV-2 COVID-19 ACE2-TMPRSS2-Furin-DPP4 Axis. *Front. Pharmacol.* 11, 584637. doi:10.3389/fphar.2020.584637

Milich, L. M., Choi, J. S., Ryan, C., Cerqueira, S. R., Benavides, S., Yahn, S. L., et al. (2021). Single-Cell Analysis of the Cellular Heterogeneity and Interactions in the Injured Mouse Spinal Cord. *J. Exp. Med.* 218 (8), e20210040. doi:10.1084/jem.20210040

Milich, L. M., Ryan, C. B., and Lee, J. K. (2019). The Origin, Fate, and Contribution of Macrophages to Spinal Cord Injury Pathology. *Acta Neuropathol.* 137 (5), 785–797. doi:10.1007/s00401-019-01992-3

Narula, S., Yusuf, S., Chong, M., Ramasundarahettige, C., Rangarajan, S., Bangdiwala, S. I., et al. (2020). Plasma ACE2 and Risk of Death or Cardiometabolic Diseases: a Case-Cohort Analysis. *The Lancet.* 396 (10256), 968–976. doi:10.1016/s0140-6736(20)31964-4

Plikus, M. V., Wang, X., Sinha, S., Forte, E., Thompson, S. M., Herzog, E. L., et al. (2021). Fibroblasts: Origins, Definitions, and Functions in Health and Disease. *Cell.* 184 (15), 3852–3872. doi:10.1016/j.cell.2021.06.024

Quinn, T. P., Richardson, M. F., Lovell, D., and Crowley, T. M. (2017). Propr: An R-Package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci. Rep.* 7 (1), 16252. doi:10.1038/s41598-017-16520-0

Raredon, M., Yang, J., Garritano, J., Wang, M., and Niklason, L. E. (2021). Connectome: Computation and Visualization of Cell-Cell Signaling Topologies in Single-Cell Systems Data. bioRxiv. doi:10.1101/2021.01.21.427529

Sha, Y., Wang, S., Bocci, F., Zhou, P., and Nie, Q. (2020). Inference of Intercellular Communications and Multilayer Gene-Regulations of Epithelial-Mesenchymal Transition From Single-Cell Transcriptomic Data. *Front. Genet.* 11, 604585. doi:10.3389/fgene.2020.604585

Shao, X., Lu, X., Liao, J., Chen, H., and Fan, X. (2020). New Avenues for Systematically Inferring Cell-Cell Communication: Through Single-Cell Transcriptomics Data. *Protein Cell.* 11 (12), 866–880. doi:10.1007/s13238-020-00727-5

Skinnider, M. A., Squair, J. W., and Foster, L. J. (2019). Evaluating Measures of Association for Single-Cell Transcriptomics. *Nat. Methods.* 16 (5), 381–386. doi:10.1038/s41592-019-0372-4

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: Guidelines and Gateway for Literature-Curated Signaling Pathway Resources. *Nat. Methods.* 13 (12), 966–967. doi:10.1038/nmeth.4077

Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., et al. (2021). Integrated Intra- and Intercellular Signaling Knowledge for Multicellular Omics Analysis. *Mol. Syst. Biol.* 17 (3), e9923. doi:10.15252/msb.20209923

Wang, Y., Wang, R., Zhang, S., Song, S., and Wang, L. (2019). iTALK: An R Package to Characterize and Illustrate Intercellular Communication. bioRxiv. doi:10.1101/507871

Zhang, L., and Nie, Q. (2021). scMC Learns Biological Variation Through the Alignment of Multiple Single-Cell Genomics Datasets. *Genome Biol.* 22 (1), 10. doi:10.1186/s13059-020-02238-2

Zhang, L., and Zhang, S. (2020). A General Joint Matrix Factorization Framework for Data Integration and its Systematic Algorithmic Exploration. *IEEE Trans. Fuzzy Syst.* 28 (9), 1971–1983. doi:10.1109/tfuzz.2019.2928518

Zhu, Y., Soderblom, C., Krishnan, V., Ashbaugh, J., Bethea, J. R., and Lee, J. K. (2015). Hematogenous Macrophage Depletion Reduces the Fibrotic Scar and Increases Axonal Growth After Spinal Cord Injury. *Neurobiol. Dis.* 74, 114–125. doi:10.1016/j.nbd.2014.10.024

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 18 January 2022
doi: 10.3389/fgene.2021.767602

MultiCapsNet: A General Framework for Data Integration and Interpretable Classification

Lifei Wang[1,2,3,4], *Xuexia Miao*[2,3], *Rui Nie*[2,3,4], *Zhang Zhang*[5], *Jiang Zhang*[5]* *and Jun Cai*[2,3,4]*

[1]Shulan (Hangzhou) Hospital Affiliated to Zhejiang Shuren University Shulan International Medical College, Hangzhou, China, [2]China National Center for Bioinformation, Beijing, China, [3]Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China, [4]University of Chinese Academy of Sciences, Beijing, China, [5]School of Systems Science, Beijing Normal University, Beijing, China

The latest progresses of experimental biology have generated a large number of data with different formats and lengths. Deep learning is an ideal tool to deal with complex datasets, but its inherent "black box" nature needs more interpretability. At the same time, traditional interpretable machine learning methods, such as linear regression or random forest, could only deal with numerical features instead of modular features often encountered in the biological field. Here, we present MultiCapsNet (https://github.com/wanglf19/MultiCapsNet), a new deep learning model built on CapsNet and scCapsNet, which possesses the merits such as easy data integration and high model interpretability. To demonstrate the ability of this model as an interpretable classifier to deal with modular inputs, we test MultiCapsNet on three datasets with different data type and application scenarios. Firstly, on the labeled variant call dataset, MultiCapsNet shows a similar classification performance with neural network model, and provides importance scores for data sources directly without an extra importance determination step required by the neural network model. The importance scores generated by these two models are highly correlated. Secondly, on single cell RNA sequence (scRNA-seq) dataset, MultiCapsNet integrates information about protein-protein interaction (PPI), and protein-DNA interaction (PDI). The classification accuracy of MultiCapsNet is comparable to the neural network and random forest model. Meanwhile, MultiCapsNet reveals how each transcription factor (TF) or PPI cluster node contributes to classification of cell type. Thirdly, we made a comparison between MultiCapsNet and SCENIC. The results show several cell type relevant TFs identified by both methods, further proving the validity and interpretability of the MultiCapsNet.

Keywords: capsule network, classification, data integration, interpretability, modular feature

OPEN ACCESS

Edited by:
Jin Chen,
University of Kentucky, United States

Reviewed by:
Md Selim,
University of Kentucky, United States
Lucas Jing Liu,
University of Kentucky, United States

***Correspondence:**
Jiang Zhang
zhangjiang@bnu.edu.cn
Jun Cai
juncai@big.ac.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *31 August 2021*
Accepted: *25 October 2021*
Published: *18 January 2022*

Citation:
Wang L, Miao X, Nie R, Zhang Z, Zhang J and Cai J (2022) MultiCapsNet: A General Framework for Data Integration and Interpretable Classification. Front. Genet. 12:767602. doi: 10.3389/fgene.2021.767602

INTRODUCTION

Recent advances in experimental biology have generated huge amounts of data. More detectable biological targets and various new measuring methods produce data at an unprecedented speed. For example, Microwell-Seq, a single cell RNA sequencing technology, has been used to analyze the transcriptome of more than 4,00,000 mouse single cells, covering all major mouse organs (Han et al., 2018); Single cell bisulfite sequencing (scBS-seq) has been designed to measure genome-wide DNA

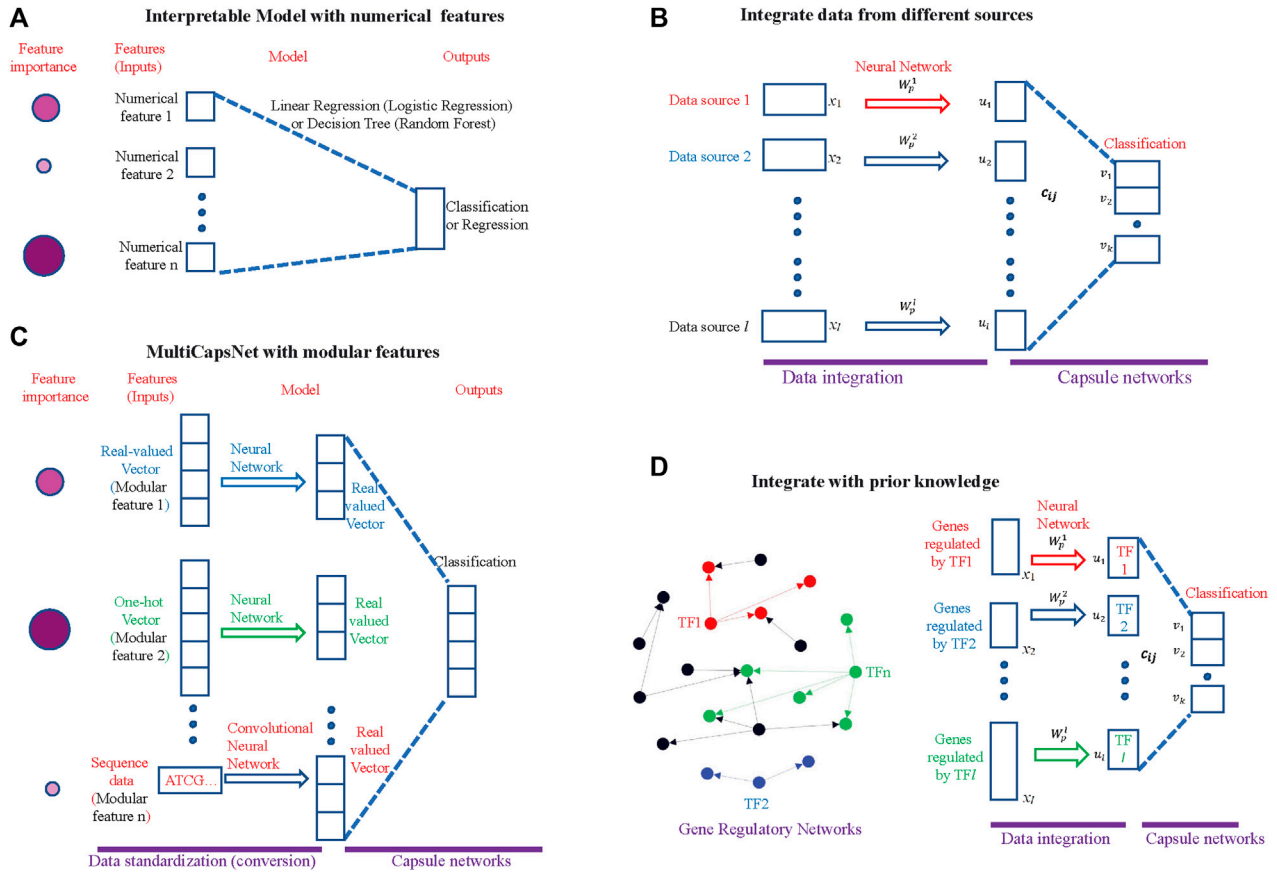


FIGURE 1 | MultiCapsNet is an interpretable classifier and data integrator with modular inputs**. (A)** The traditional interpretable machine learning methods. The input of this model is numerical. After training, the model will reveal the inputs (features) importance for classification (or regression). The size and color depth of the circle indicate the importance of the features, while the larger and darker circle indicates that the feature is more important. **(B)** The MultiCapsNet is an interpretable classifier with modular input. The inputs (features) with different format (real-valued vector, one-hot encoding vector, or sequence data) and different lengths are first converted into real-valued vectors with equal length through trainable networks. Then, classification was based on those real-valued vectors of the same length. After training, the model will reveal the inputs (features) importance to classification. The size and color depth of the circle indicate the importance of the feature, while larger and darker circle indicates that the feature is more important. **(C)** The MultiCapsNet could integrate data from different sources. **(D)** The MultiCapsNet could integrate prior knowledge, such as gene regulatory information. Left: Gene regulatory networks, transcription factor and its targets are marked with same color. Right: expression of genes that are regulated by the same transcription factor could be regarded as a data source.

methylation at the single-cell level (Smallwood et al., 2014); and mass-spectrometry based technologies could explore the composition, structure, function, and control of the proteome (Aebersold and Mann, 2016). In addition, large and complex data sets are produced by large-scale projects, such as "The Cancer Genome Atlas" (TCGA) (Tomczak et al., 2015), and "Encyclopedia of DNA Elements" (ENCODE) (Consortium, 2004), which were established through community cooperation. There is an urgent need for next generation methods to deal with large, heterogeneous and complex data sets (Camacho et al., 2018).

As a promising data processing method, deep learning methods have been employed in biological data processing (Alipanahi et al., 2015; Camacho et al., 2018; Zhou et al., 2018; Eraslan et al., 2019). Various deep learning models could deal with various input data with different types and formats. For example, RNA sequence data as real-value vectors could be processed by simple feed forward neural network, which is a component of more complex models, such as auto-encoder (AE) (Lin et al., 2017; Chen et al., 2018), variational auto-encoder (VAE) (Ding et al., 2018), and Generative adversarial network (GAN) (Lopez et al., 2018). Sequence information, which is coded by ATCG, could be converted into real valued vectors by deep learning model using convolution neural networks (CNN) after model training (Alipanahi et al., 2015). Furthermore, deep learning models could integrate data with different types and formats. For example, DeepCpG utilizes both DNA sequence patterns and neighboring methylation states for predicting single-cell methylation state and modeling the sources of DNA methylation variability (Angermueller et al., 2017). However, the deep learning methods usually run as a "black box", which is hard to interpret (Almas Jabeen and Raza, 2017). Great efforts have been made to improve the interpretability of deep learning models. The prior biological information, such as regulation

between transcription factors (TF) and target genes or priori defined gene sets that retain the crucial biological features, could specify connections between neurons in the neural networks in order to associate the internal node (neuron) in the neural networks with TFs and thereby ease the difficulty of interpreting models (Lin et al., 2017; Chen et al., 2018). New probabilistic generative models with more interpretability, such as variational inference neural networks, are applied to scRNA-seq data for dimension reduction (Ding et al., 2018).

Traditional interpretable machine learning methods, such as linear regression (logistic regression) or decision tree (random forest), could only deal with numerical or categorical feature (Molnar, 2019) (**Figure 1A**). However, in the field of biology, especially in the field of network biology, the data is highly modular in nature. For example, in drug discovery, many independent features with multiple labels (e.g., response to drug, and disease state) across a multitude of data types (e.g., expression profiles, chemical structures) are needed; and in synthetic biology, the input may include sequence data, composition data and functional data (Camacho et al., 2018). An interpretable machine learning method adapted with modular input is demanded.

The capsule network (CapsNet) is a newly developed deep learning model for digital recognition tasks (Sabour et al., 2017). In the realm of biology, the CapsNet model has been directly applied for protein structure classification and prediction (Dan Rosa de Jesus et al., 2018; Fang et al., 2018) and is ripe for application in network biology and disease biology with data from multi-omics dataset (Camacho et al., 2018). In our previous work, we proposed a modified CapsNet model, called single cell capsule network (scCapsNet), which is suitable for single-cell RNA sequencing (scRNA-seq) data (Wang et al., 2020). The scCapsNet is a highly interpretable cell type classifier, with the capability of revealing cell type associated genes by model internal parameters.

Here, we introduce MultiCapsNet, a deep learning classifier and data integrator built on CapsNet and scCapsNet. As a general framework, the MultiCapsNet model should be able to deal with modular data from multiple sources with different formats and lengths, and give the importance scores of each data source for prediction after training (**Figures 1B–D**). In order to demonstrate its wide biological application, the MultiCapsNet model was tested on three data sets. In the first example, we applied the MultiCapsNet model to the labeled variant call data set, which was originally used to test the models for automating somatic variant refinement (Ainscough et al., 2018). According to data source and data attributes, the 71 features listed in the data set were divided into eight groups. Then the features in one group were viewed as a whole to train the MultiCapsNet model. After training, the performance of our MultiCapsNet matches well with the previous feed forward neural network model and random forest model. As an advantage our MultiCapsNet model directly provides the importance score for each data source, while the previous feed forward neutral network model needs an extra importance determination step through shuffling individual features to do so. Despite that our MultiCapsNet model is substantially different from the previous feed forward neural network model and the source importance measuring methods are also different, the correlation between the importance scores generated by those two models is highly correlated. In the second example, we demonstrate how to integrate prior knowledge and scRNA-seq data through MultiCapsNet model. The protein-protein interactions (PPI) information stored in BIOGRID (Stark et al., 2006) and HPRD (Keshava Prasad et al., 2009), and protein-DNA interactions (PDI) from DREM 2.0 (Schulz et al., 2012), are used as prior knowledge to specify network connections, as in previous work (Lin et al., 2017). In this example, the structures of the first part of the MultiCapsNet model, i.e., the connections between inputs and primary capsules, are determined by the PPI and PDI information. As a result of these specified structures, each primary capsule is labeled either as TF or PPI subnetwork (PPI), and inputs of each primary capsule could be regarded as a data source. We use data from mouse scRNA-seq dataset (Han et al., 2018) to train this MultiCapsNet model and the classification accuracy of MultiCapsNet is comparable to neural network and random forest model. After training, the MultiCapsNet model reveals how each primary capsule, which is labeled either as TF or PPI subnetwork (PPI), contributes to cell type classification. The top contributors of a particular cell type are usually related to that cell type. In the third example, we make a comparison between our MultiCapsNet and the established single-cell regulatory network inference method: SCENIC (Single-cell regulatory network inference and clustering) (Aibar et al., 2017). The results show that many cell types relevant TFs are identified by both methods, which further proves the validity and interpretability of MultiCapsNet.

METHODS

Datasets and Data Preprocessing

Labeled variant call dataset from previous work was used to test the MultiCapsNet model (Ainscough et al., 2018). This dataset contains more than 41,000 samples, which are assembled to train models for automating somatic variant refinement. Each sample in the dataset is manually labeled as one of four tags by the reviewer: “somatic”, “ambiguous’, “germline”, and “fail”, which represent the confidence of a variant call by upstream somatic variant caller. As in previous work, we merged the variant calls labeled as “germline” and “fail” into a class named “fail”. The number of instances in each class are around 10,000, 13,000, 18,000 for “ambiguous”, “fail”, and “somatic”. There are 71 features that are associated with each sample, including cancer types, reviewers, tumor read depth, normal read depth, and so on. According to the data sources and data attributes, we divided these 71 features into eight groups (**Supplementary Table S1**). Group 1 contains nine cancer types, and is encoded as one-hot encoding vector. We call group 1 as “Disease” because it indicates the disease to which each variant call belongs. Group 2 contains four reviewers, and is encoded as one-hot encoding vector. We call group 2 as “Reviewer”. Group 3 contains information of “normal VAF”, “normal depth”, “normal other bases count”, and is called as “Normal_pro”, short for “Normal properties”. Group

4 contains 13 features that describe reference reads in normal, including base quality, mapping quality, numbers of mismatches, numbers of minus and plus strand, and so on. We call group 4 as "Normal_ref". Group 5 contains 13 features extracted from variant reads in normal, also including base quality, mapping quality, numbers of mismatches, numbers of minus and plus strand, and so on. We call group 5 as "Normal_var". The last three groups contain features drawn from tumor instead of normal in previous three groups. As same as Group 3, 4, and 5, we label group 6, 7, and 8 as "Tumor_pro", "Tumor_ref", and "Tumor_var" respectively.

The mouse scRNA-seq is measured by Microwell-Seq (Han et al., 2018). We downloaded scRNA-seq data and the annotation information through the link provided by the authors (https://figshare.com/s/865e694ad06d5857db4b). Then we use the annotation information to select parts of data from whole dataset. The cell types we chose include "Cartilage cell", "Secretory alveoli cell", " Epithelial cell ", "Kupffer cell", "Muscle cell", "Dendritic cell", " Spermatocyte", and the number of instances in each cell type are 527, 1,195, 1,219, 356, 626, 717, 353. Moreover, we only use the genes contained in prior knowledge (Lin et al., 2017) to fit the model structure, and set the default value to zero when the downloaded scRNA-seq data does not contain this gene (Han et al., 2018).

A SCENIC example dataset was used to compare the performances of MultiCapsNet and SCENIC (https://scenic.aertslab.org/examples/). The dataset (sceMouseBrain.RData) contains seven cell types of mouse cortex and hippocampus (Zeisel et al., 2015) ["astrocytes_ependymal" (224), "endothelial_mural" (235), "interneurons" (290), "microglia" (98), "oligodendrocytes" (820), "pyramidal_CA1" (939), and "pyramidal_SS" (399)].

The Architecture and Parameters of the MultiCapsNet Model

In the architecture of our multiCapsNet model, there are l neural networks corresponding to l input modular data.

$$u_i = tanh\left(W_p^i x_i\right) \quad i \in [1, 2 \ldots, l] \tag{1}$$

x_i represents i's input modular data. W_p^i represents weight matrices of neural networks with dimension (n, r_i), where the r_i is the length of the input modular data x_i. The output u_i of each neural network i ($i \in [1, 2 \ldots, l]$) is a vector with length n, viewed as "primary capsule" in the model. The inputs standardization part converts the modular data with different type and length into real valued vectors with equal length n ($n = 8$ by default).

The standardized information is subsequently delivered through primary capsule to the capsule in the final layer by "dynamic routing" (**Supplementary Figure S1**). Each capsule in the final layer, named "type capsule", corresponds to each cell type. They are denoted as vectors v_j, where $j \in [1, 2 \ldots, k]$, k is the number of cell types and m is the length of vectors. The capsule network module is implemented in Keras (https://github.com/bojone/Capsule).

Prior to the "dynamic routing" process, the primary capsules are multiplied by weight matrices W_{ij} to produce "prediction vectors" $\hat{u}_{j|i}$.

$$\hat{u}_{j|_i} = W_{ij} u_i \tag{2}$$

Then the iterative dynamic routing begins. Firstly, the "coupling coefficients" $c_{ij}$is calculated by formula:

$$c_{ij} = \frac{\exp\left(b_{ij}\right)}{\sum_k \exp\left(b_{ik}\right)} \tag{3}$$

Where b_{ij} is an intermediate parameter with initial value of zero, representing the inner product of the prediction vector and type capsule vector.

In order to compute the b_{ij} for next round iteration, the weighted sum s_j over all k prediction vectors $\hat{u}_{j|i}$ is calculated by formula:

$$s_j = normalize\left(\sum_i c_{ij} \hat{u}_{j|_i}\right) \tag{4}$$

Secondly b_{ij} is computed by the dot product of $\hat{u}_{j|i}$ and s_j as the last step of one round dynamic routing process.

$$b_{ij} = \hat{u}_{j|_i} . s_j \tag{5}$$

After several rounds of dynamic routing, the type capsule v_j is calculated by a non-linear "squashing" function:

$$v_j = \frac{\left\|s_j\right\|^2}{0.5 + \left\|s_j\right\|^2} \frac{s_j}{\left\|s_j\right\|} \tag{6}$$

The following pseudocode illustrates the implementation of MultiCapsNet.

1) *for all primary capsule i*: $u_i = Activation\ Function\left(W_p^i x_i\right)$
2) for *all primary capsule i* and type capsule *j*: $\hat{u}_{j|i} = W_{ij} u_i$
3) procedure ROUTING$\left(\hat{u}_{j|i}, r\right)$
4) for all primary capsule I and type capsule *j*: $b_{ij} \leftarrow 0$.
5) For r iterations do
6) for all primary capsule *i*: $c_i \leftarrow$ softmax (b_i)
7) for all type capsule *j*: $s_j \leftarrow normalize\left(\sum c_{ij} \hat{u}_{j|i}\right)$
8) for all primary capsule *i* and type capsule *j*: $b_{ij} \leftarrow \hat{u}_{j|i} . s_j$

return $v_j \leftarrow$ squash (s_j)

The implementation of MultiCapsNet can be found in https://github.com/wanglf19/MultiCapsNet.

MultiCapsNet Model in Somatic Variant Refinement Task

In the somatic variant refinement task, the eight groups mentioned above in the section of "Datasets and data preprocessing" correspond to eight input sources. Therefore, there are eight neural networks corresponding to eight groups of input modular data ($l = 8$).

$$u_i = tanh\left(W_p^i x_i\right) \quad i \in [1, 2 \ldots, 8] \tag{7}$$

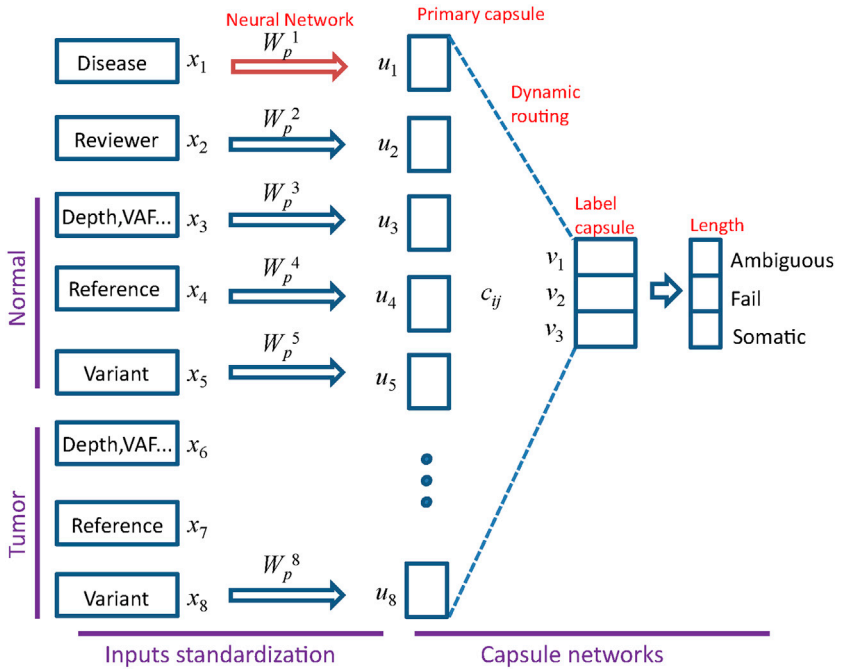


FIGURE 2 | Architecture of MultiCapsNet with two layers. The first layer consists of eight parallel neural networks, corresponding to eight data sources (groups). The outputs of neural networks are the primary capsules (real valued vectors) with equal length. The second layer is the Keras implementation of CapsNet for classification. The length of each label capsule represents the probability that the input data belongs to the corresponding classification category.

After the input standardization part, the input data x_i is converted into a primary capsule u_i having the same length. Next, the standardized information stored in the primary capsules would be delivered to the final layer capsules by "dynamic routing". The capsules in the final layer, which corresponds to labels of variant calls, is called "label capsule". In capsnet, the non-linear "squashing" function ensure that short vectors get shrunk to almost zero length and long vectors get shrunk to a length slightly below 1 (Sabour et al., 2017). The length of the label capsule represents the probability that a variant call is either "ambiguous","fail", or "somatic" (**Figure 2**). To evaluate the performance of the model, we use the "area under the curve" (AUC) score as previous (Ainscough et al., 2018) and prediction accuracy.

MultiCapsNet Model That Integrates Prior Knowledge

The MultiCapsNet could integrate prior knowledge into its structure. In brief, PPI information store in BIOGRID (Stark et al., 2006) and HPRD (Keshava Prasad et al., 2009), and PDI coming from DREM 2.0 (Schulz et al., 2012), are used as prior knowledge for specifying network connections between the inputs and the primary capsules (**Figure 4A**), just as previous work used this prior knowledge to specify network connections between the inputs and neurons (Lin et al., 2017). For example, the prior knowledge indicates that $Gene_1$,. . ., $Gene_n$ are regulated by a TF (colored with green), so there are connections between $Gene_1$,. . ., $Gene_n$ and primary capsule representing corresponding TF (green connection); the prior knowledge indicate that $Gene_2$,. . ., $Gene_n$ are regulated by a TF (colored with blue), then there are connections between $Gene_2$,. . ., $Gene_n$ and primary capsule representing corresponding TF (blue connection); and the prior knowledge indicates that $Gene_2$, $Gene_3$,. . ., are in a subnetwork of PPI network (colored with red), then there are connections between $Gene_2$, $Gene_3$,. . ., and primary capsule representing corresponding PPI subnetwork (red connection). Although there is only one input source, namely scRNA-seq data, the input source can be decomposed into several parts by integrating prior knowledge, and each part is connected to a primary capsule. Therefore, we also took a single input source integrated with prior knowledge as an input from multiple sources, each of which is associated with a TF or a PPI subnetwork (**Figure 4B**).

In total there are 696 input modular data, with 348 TF-targets relationships extracted from PDI information and 348 PPI subnetworks. Therefore, there are 696 neural networks corresponding to 696 modular data ($l = 696$).

$$u_i = tanh\left(W_p^i x_i\right) \quad i \in [1, 2 \ldots, 696] \tag{8}$$

After the input standardization part, the input data x_i is converted into primary capsule u_i with the same length. Next, the standardized information stored in the primary capsules would be delivered to the final layer capsules by "dynamic routing". The capsules in the final layer, which correspond to cell types, is called "type capsule".

MultiCapsNet Model Compared with SCENIC

The SCENIC is a workflow for simultaneous reconstruction of gene regulatory networks and identification of cell states using scRNA-seq data (Aibar et al., 2017). The workflow consists of three

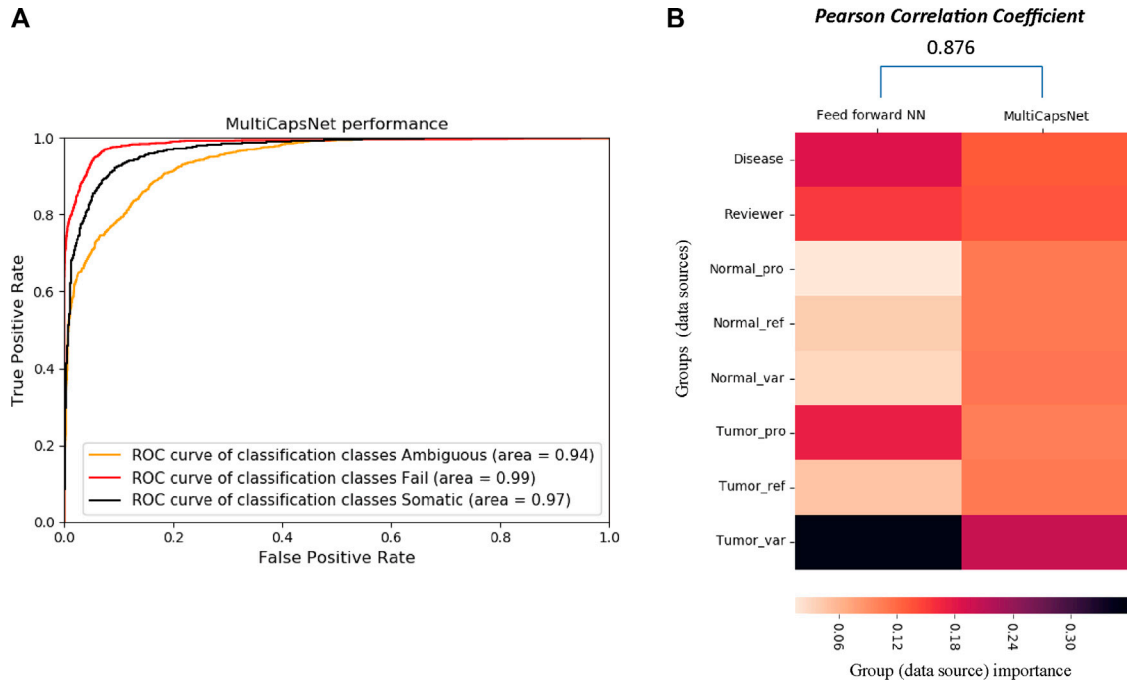


FIGURE 3 | Architecture of MultiCapsNet integrated with prior knowledge. **(A)** The model has two layers. The first layer consists of 696 parallel neural networks corresponding to 696 primary capsules labeled with either transcription factor (348) or protein-protein interaction cluster node (348). The inputs of each primary capsule include genes regulated by a transcription factor or in a protein-protein interactions sub-network. The second layer is the Keras implementation of CapsNet for classification. The length of each final layer type capsule represents the probability of input data belonging to the corresponding classification category. **(B)** Alternative representation of MultiCapsNet integrated with prior knowledge. Genes that are regulated by a transcription factor or in a protein-protein interactions sub-network, are groups together as a data source for MultiCapsNet. **Figures 3A,B** are equivalent with different representation.

modules (R/bioconductor packages): GENIE3 (GRNboost), RcisTarget, AUCell. The first two modules were responsible to find potential TF-targets relationships based on co-expression and subsequently select the highly confident TF-target regulation according to TF-motif enrichment analysis. After that, several potential TF-target relationships across all cell types, called regulons, were identified in the dataset. The AUCell would score the activity of these regulons in each single cell. Finally, the unsupervised method is used to cluster cell, identify cell types and states based on the scores of the regulongs, which are used as features for each cell. In our model, we utilized the regulon information identified by the first two modules of SCENIC as the prior knowledge to specify the connections between input and primary capsules (**Supplementary Figure S2A**). The dataset, intermediate results and the output of SCENIC for a mouse brain example were downloaded from the website (https://scenic.aertslab.org/examples/). The regulon information was extracted from the intermediate result file (regulons_asGeneSet.Rds).

In total there are 253 regulons, which specify TFs and their target genes. Therefore, there are 253 neural networks corresponding to 253 modular data ($l = 253$).

$$u_i = tanh\left(W_p^i x_i\right) \quad i \in [1, 2 \ldots, 253] \tag{9}$$

After the input standardization part, the input data x_i is converted into primary capsule u_i with same length. Next, the standardized information stored in the primary capsules would be delivered to the final layer capsules by "dynamic routing". The capsules in the final layer, which correspond to cell types, is called "type capsule".

Average Coupling Coefficients and Data Source Importance

In scCapsNet, we showed that the average coupling coefficients represent the contribution of the primary capsules to the final layer type capsules for each cell type (Wang et al., 2020). Similarly, in the multiCapsNet model, the type (label) capsule v_j derives from a weighted sum of prediction vectors $\hat{u}_{j|i}$. The weights are the coupling coefficients c_{ij} and the magnitude of these coefficients could roughly be regarded as the contribution of the primary capsules u_i to the type capsules v_j. Each sample (single cell, somatic variant) generates its own coupling coefficients. The average coupling coefficients for samples with same type (label) are calculated by the formular:

$$c_{ij}^{type\ average} = \frac{\sum_{type} c_{ij}^{type}}{\sum_{type} 1} \tag{10}$$

Therefore, each classification category (cell type/variant call label) corresponds to an average coupling coefficients matrix ($c_{ij}^{type\ average}$), called type average coupling coefficients, with rows representing type capsules and columns representing primary capsules. The type average coupling coefficients matrix could be plotted as heatmap for visualization of data. For each classification

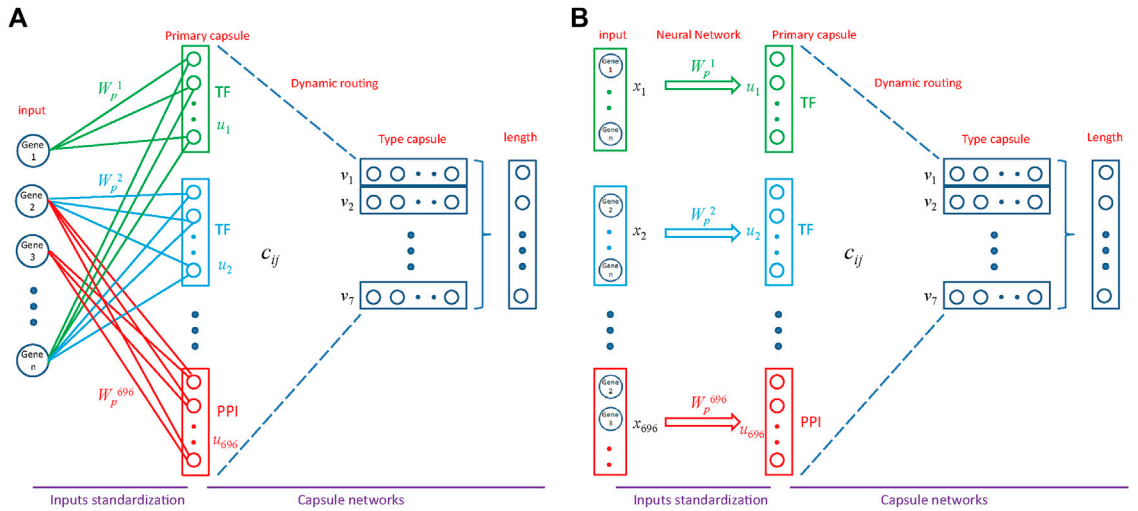


FIGURE 4 | The comparison between MultiCapsNet and feed forward neural network shows the high performance and interpretability of MultiCapsNet. **(A)** The AUC scores demonstrate that the MultiCapsNet model achieves very high classification performances in all three classification categories. **(B)** The normalized group (data source) importance scores generated by MultiCapsNet and feed forward neural network are highly correlated.

category (cell type/variant call label), the corresponding type average coupling coefficients matrix contain an effective type capsule row, which is the row whose type is consistent with this classification category. For example, the effective type capsule row in the type average coupling coefficients matrix ($c_{ij}^{type\ 2\ average}$) is the row $c_{i2}^{type\ 2\ average}$. In this row, the magnitude of each element could be regarded as the importance score of the corresponding primary capsule to this classification category. The effective type capsule rows of all classification categories ($c_{i1}^{type\ 1\ average}$, $c_{i2}^{type\ 2\ average}$, $c_{i3}^{type\ 3\ average}$...) could be organized into a new matrix, visually represented as an overall heatmap.

Algorithm Implementation for Comparisons

A neural network with sigmoid activation function was implemented in Keras. The random forest and nearest-neighbour are implemented with the Python package "scikit-learn". The comparison transformers model was originally used for IMDB movie review sentiment classification dataset. This transformer model contains the embedding layer for embedding the words into vectors and the Multi-head attention layer (https://github.com/bojone/attention/). We replace the embedding layer with our data standardization layer, and retain the Multi-head attention layer for classification.

RESULTS

MultiCapsNet Achieves High Classification Accuracy and High Interpretability for Modular Data From Variant Call Dataset

The variant call dataset (Please refer to Datasets and data preprocessing in METHODS section for the details) was randomly divided into training set and validation set with a ratio of 9:1. Our MultiCapsNet model performs well in the classification of variant call (**Figure 4**). The results show that the AUC of the MultiCapsNet model is 0.94, 0.99, and 0.97, respectively, in the classification categories of "ambiguous", "fail", and "somatic" (**Figure 4A**). These AUC scores are similar with those obtained by the Multi-head Attention model (0.93, 0.98, 0.96), feed forward neural network (0.93, 0.99, 0.96), and random forest (0.96, 0.99, 0.98) (Ainscough et al., 2018). Meanwhile, the average prediction accuracy of the MultiCapsNet model is around 0.873, similar to those obtained by the Multi-head Attention model (0.866), and slightly lower than that of feed forward neural network (0.887), and random forest (0.895).

In MultiCapsNet, the coupling coefficient c_{ij} is viewed as important scores, which is the weight that measure the contribution of each primary capsule to the final layer type capsule. Each input would generate its own coupling coefficient, and the type average coupling coefficient is the average over all the inputs with same classification category. After MultiCapsNet model training, the type average coupling coefficients for each variant label ("ambiguous", "fail", and "somatic") were calculated and visualized as heatmaps (**Supplementary Figure S3A**) (Please refer to "METHODS" section for the detailed calculation formula of type average coupling coefficients). In each type average coupling coefficient, the most important row, named as "effective type capsule row", is the row whose type is consistent with this classification category. The overall heatmap is assembled with the "effective type capsule row" which describes the importance scores of all the data sources for distinct category classification (**Supplementary Figure S3B**). Therefore, the overall heatmap also shows the contribution of each data source to the recognition of each variant labels ("ambiguous", "fail", and "somatic"). For example, the data source of "Disease" has the contribution to the classification of "somatic" category and the "Reviewer" source contributes to the classification of "ambiguous"

category. The "Tumor_var" source is the most important one for the classification of all the three categories (**Supplementary Figure S3B**). Over 9 repetitions, the values of each row in 9 overall heatmap are averaged to determine the importance scores of each data sources for the classification of all the categories in MultiCapsNet model (**Figure 3B**). In feed forward neural network model, the feature importance is measured by average change of AUC after randomly shuffling individual features. Based on the step of features grouping, we added the feature importance scores belonging to the same group together, and take these values as importance of data sources (each group) in feed forward neural networks model (**Figure 3B**). Then, we calculated the correlation between the data source importance scores obtained by our MultiCapsNet model and those provided by feed forward neural network model. Although our MultiCapsNet model is substantially different from the previous feed forward neural network, and the source importance measuring methods are also different, there is very high correlation between them (Pearson Correlation Coefficient = 0.876) (**Figure 4B**). Both models indicate that tumor variant group is very important for variant call classification.

MultiCapsNet Integrated with Prior Knowledge Could Function as Classifier and Identify Cell Type Relevant TF

The dataset is a portion of mouse scRNA-seq data measured by Microwell-Seq, which consists of nearly 5,000 cells of seven types and 9,437 genes (Please refer to METHODS section for the details). The MultiCapsNet model that integrates prior knowledge (**Figure 4**) was trained and tested by using this dataset. The average validation accuracy and F1 score are around 97%, comparable with those generated by the feed forward neural network, Multi-head Attention model and random forest (**Supplementary Figures S4A, B**). After training, the average coupling coefficients, which represent the contribution of the primary capsules (TF/PPI) to the type capsules (Cell type), were calculated and visualized as heatmaps for each cell type (**Figure 5A**). In each heatmaps, we should clearly observe that the high value elements in the average coupling coefficients (dark line in the plot) are exclusively located in the effective type capsule row. Then, the corresponding type capsule row was selected from each heat map in **Figure 5A**, and organized into an overall heatmap (**Figure 5B**).

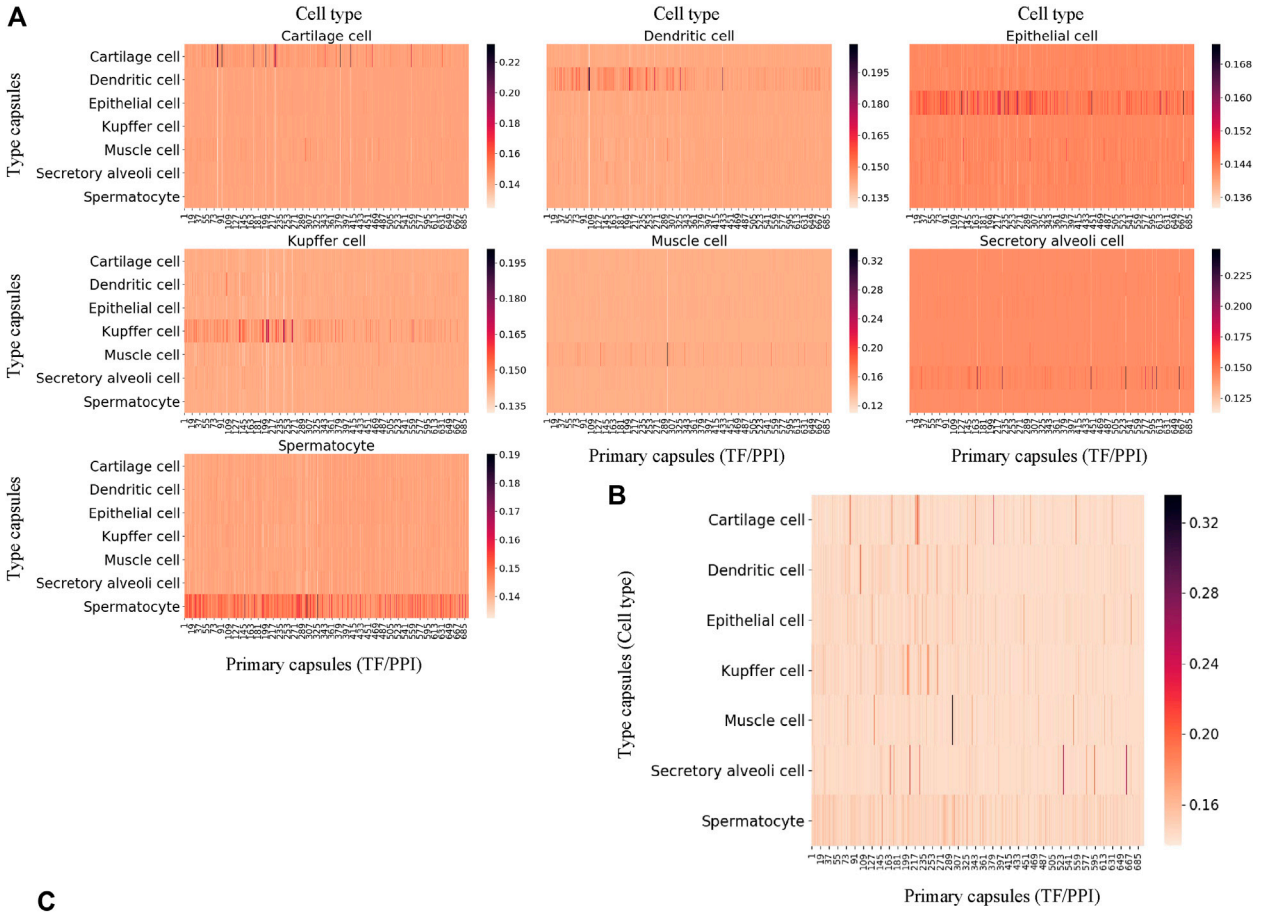
We repeat the training process 9 times and generate nine overall heatmaps accordingly. Based on the average value of the nine overall heatmaps, the top 10 relevant TFs/PPI subnetwork was generated (**Supplementary Table S2**). Most of the top 10 relevant TFs/PPI subnetwork were specific to one cell type, and many of them have been reported to be associated with corresponding cell types previously (**Figure 5C**). For example, *Gata1* and *Gata2* are top contributors for dendritic cell recognition. Previous work indicated that *Gata1* regulates dendritic cell development and survival (Gutiérrez et al., 2007), *Gata2* regulates dendritic cell differentiation (Onodera et al., 2016). *Srf* and *Yy1* are ranked as the top contributors for muscle cell recognition by the model. However, *Srf* is required for skeletal muscle growth and maturation (Li et al., 2005), *Yy1* is associated with increased smooth muscle specific gene expression (Favot et al., 2005). *FoxA2* and *FoxA3* are ranked as top contributors for Cartilage cell recognition, and *FoxA2* and *FoxA3* are necessary to promote high-level expression of several hypertrophic chondrocyte markers (Ionescu et al., 2012). The model reports *Rxrg*, *Rara*, *Rarg*, *Rarb*, *Rxra*, and *Rxrb* as top contributors for Kupffer cell recognition. Previous research report RA receptor (RAR) and retinoid X receptor (RXR) were expressed by Kupffer cells (Ulven et al., 1998; Ohata et al., 2000). *Pgr* is ranked as a top contributor for secretory alveoli cell recognition, and the progesterone receptor (*Pgr*) knockout mouse demonstrated that Pg is required for alveolar morphogenesis (Oakes et al., 2006). *Topors* is ranked as a top contributor for spermatocyte recognition. Previous work indicates *dtopors*, the *Drosophila* homolog of the mammalian *Topors*, plays a structural role in spermatocyte lamina that is critical for multiple aspects of meiotic chromosome transmission (Matsui et al., 2011).

The Comparison of MultiCapsNet Model with SCENIC Shows That Several Cell Type Relevant TFs Are Identified by Both Methods

To further demonstrate the effectiveness of our MultiCapsNet model to reveal cell type related TFs from scRNA-seq data, we compare it with established single-cell regulatory network inference methods: SCENIC (Single-cell regulatory network inference and clustering) (**Supplementary Figure S2A**). The scRNA-seq data from mouse cortex and hippocampus were used to evaluate these two methods (Please refer to METHODS section for the details).

After MultiCapsNet training, the average coupling coefficients in the overall heatmap would indicate the most relevant TFs associated with each cell type (**Supplementary Figure S5**). We repeated the experiment 9 times, the average validation accuracy was 97%, and the average F1 score was around 95%, which were comparable to the results generated by feed forward neural network, Multi-head Attention model and random forest (**Supplementary Figures S4C, D**). According to the average value of nine overall heatmaps, the top 30 relevant TFs could be generated (**Figure 6A** left; **Supplementary Table S3** top). The original regulon may contain TFs that label the 253 regulons. In order to eliminate the influence caused by the expression of those labeling TF, the potential TF-target relationships that exclude the labeling TF in the set of target genes are also made (**Supplementary Figure S2B**). We also repeated the training process of MultiCapsNet that integrated with those new potential TF-target relationships. After training, the top 30 relevant TFs could also be generated according to the average value of the nine overall heatmaps (**Figure 6A** right; **Supplementary Table S3** bottom). The results show that the inclusion or exclusion of labeling TF has little influence on prediction accuracy and interpretability of the model. The overlap rates of top 30 most relevant TF of each cell type (around top 10% of total TFs) between model including labeling TF and that excluding labeling TF are very high, around 90% for every cell type (**Figure 6B**).

Many high score TFs predicted by MultiCapsNet are consistent with that reported by SCENIC (Aibar et al., 2017). For example, in both methods, *Rorb* is identified as a relevant TF

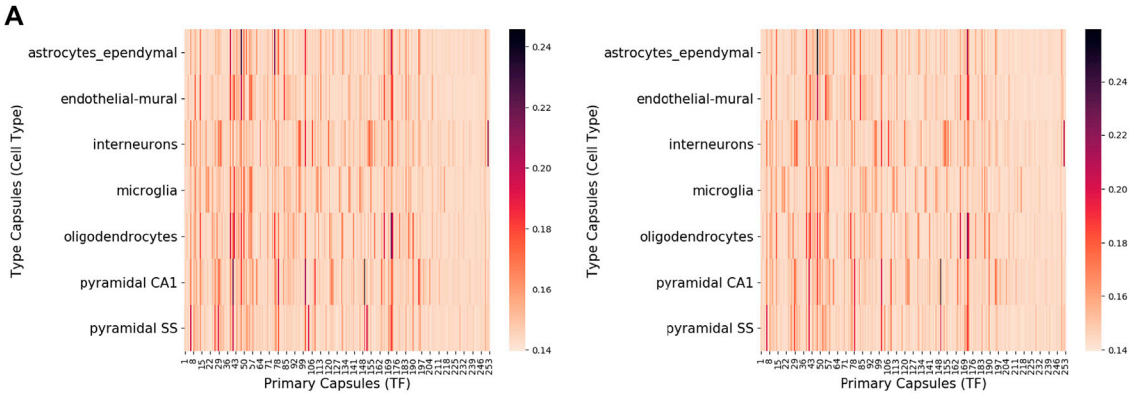


C

Cell type	TF	Previous researchs
Dendritic cell	*Gata1*	*Gata1* regulates dendritic-cell development and survival.
Dendritic cell	*Gata2*	*Gata2* regulates dendritic cell differentiation.
Muscle cell	*Srf*	Requirement for serum response factor(*Srf*) for skeletal muscle growth and maturation.
Muscle cell	*Yy1*	Cytoplasmic *Yy1* Is Associated with Increased Smooth Muscle-Specific Gene Expression.
Cartilage cell	*FoxA2 and FoxA3*	*FoxA2* and *FoxA3* are necessary to promote high-level expression of several hypertrophic chondrocyte markers
Kupffer cells	*Rxrg, Rara, Rarg, Rarb, Rxra ,Rxrb*	RA receptor (RAR) and retinoid X receptor (RXR) express by Kupffer cells
Secretory alveoli cell	*Pgr*	the progesterone receptor (*Pgr*) knockout mouse demonstrated that Pg is required for alveolar morphogenesis.
Spermatocyte	*Topors*	*dtopors*, the Drosophila homolog of the mammalian *Topors*, plays a structural role in spermatocyte lamina that is critical for multiple aspects of meiotic chromosome transmission.

FIGURE 5 | MultiCapsNet integrated with prior knowledge could identify cell type associated transcription factor. **(A)** heatmaps of the matrices of average coupling coefficients for each cell type. In each heatmap, there are 696 columns for 696 primary capsules (TF/PPI) and seven row for seven type capsules (cell types), and each element in the average coupling coefficients is represented by a thin line. The brightness of these thin lines (elements in the average coupling coefficients) indicate the contribution of the primary capsules (TF/PPI) to the specific cell type recognition. The dark lines (high score elements in average coupling coefficients) exclusively reside in the corresponding effective type capsule row in each heatmap. **(B)** Overall heatmap of the combined matrix of average coupling coefficients. The combined matrix contains the effective type capsule rows in **Figure 5A** where its recognition type is in accordance with the type of single cells input. **(C)** The table list several top ranked contributors for specific cell type recognition, given by the MuiltCapsNet model, are associated with corresponding cell types which have been reported before.

for astrocytes; *Ets1*, *Elk3*, and *Gata2* are identified as relevant TFs for endothelial-mural cells; *Zmat4*, *Dlx5*, *Dlx2*, and *Dlx1* are identified as relevant TFs for interneurons; *Maf*, *Rel*, *Cebpa*, *Cebpb*, *Nfatc2*, *Prdm1*, *Nfkb1*, and *Stat6* are identified as relevant TFs for microglia; *Sox10* and *Sox8* are identified as relevant TFs for oligodendrocytes. Besides the TFs listed above, MultiCapsNet also detected several high confidence cell type relevant TFs that are also found by SCENIC. For example, *Rfx3* shows a high association with both pyramidal SS and CA1 cells. Previous studies reported that downstream target of *Rfx3*



B

cell type	astrocytes_ependymal	endothelial-mural	interneurons	microglia	oligodendrocytes	pyramidal CA1	pyramidal SS
overlap(%)	90(27/30)	90(27/30)	86.7(26/30)	86.7(26/30)	96.7(29/30)	86.7(26/30)	90(27/30)

C

Cell type	TF
astrocytes_ependymal	*Rorb*
endothelial-mural	*Ets1, Elk3, Gata2*
interneurons	*Zmat4, Dlx5, Dlx2, Dlx1*
microglia	*Maf, Rel, Cebpa, Cebpb, Nfatc2, Prdm1, Nfkb1, Stat6*
oligodendrocytes	*Sox10, Sox8, Ezh2*
pyramidal CA1	*Rfx3,Neurod2,smarca4*
pyramidal SS	*Rfx3,Neurod2,Cux1*

FIGURE 6 | The comparison of MultiCapsNet and SCENIC shows the robustness and interpretability of MultiCapsNet. **(A)** Averaged overall heatmaps for mouse cortex and hippocampus dataset show that MultiCapsNet perform consistently whether including**(left)** or excluding **(right)** the labelling TF from regulon. **(B)** The top ranked contributors for specific cell type classification identified from dataset either including **(left)** or excluding **(right)** the labelling TF are highly overlapped. **(C)** The table list several top ranked contributors for specific cell type recognition, given by both the MuiltCapsNet model and SCENIC.

displayed cytosolic expression in pyramidal neurons (Remnestål, 2015) and *Rfx3* expresses in cortical pyramidal neurons (Benadiba et al., 2012). *Neurod2* is also identified as a relevant TF for both pyramidal SS and CA1 cells. Previous studies reported that *Neurod2* coordinates synaptic innervation and cell intrinsic properties to control excitability of cortical pyramidal neurons (Chen et al., 2016). *Cux1* has been identified as a relevant TF for pyramidal SS cells, and *Cux1* has been reported as a restricted molecular marker for the upper layer (II-IV) pyramidal neurons in murine cerebral cortex (Li et al., 2010). *smarca4* has been identified as relevant TF for pyramidal CA1 cells, and *Brg1/smarca4* deficiency leads to mouse pyramidal neuron degeneration (Deng et al., 2015). *Ezh2* has been suggested as a relevant TF for oligodendrocytes, and the expression of *Ezh2* in OPCs (oligodendrocytes precursor cells), even up to the stage of pre-myelinating immature oligodendrocytes, remains high (Copray et al., 2009) (**Figure 6C**). Furthermore, the MultiCapsNet found that *Rpp25* is strongly associated with interneurons which SCENIC did not, and *Rpp25* has been reported up-regulated in GABAergic interneuron (Fukumoto et al., 2018).

DISCUSSION

In the first example, we demonstrated that the proposed MultiCapsNet model performed well in the variant call classification. Data sources with different data types, such as one-hot encoding vector and real valued vectors, could be standardized into equal length vectors as primary capsules, and then pass the information into final layer capsules by dynamic routing. The importance of the data sources was measured by the sum of the overall average coupling coefficients as the co-product of the model training. These importance scores are highly correlated with the importance scores calculated by feed forward neural network, which are measured by average change in the AUC after randomly shuffling individual features.

In the second example, we incorporated PPI and PDI information into the structure of the MultiCapsNet model. This specified structure

decomposed the input scRNA-seq data into several parts, each part corresponding to a group of genes regulated by a TF or from a protein interaction sub-network. Therefore, each part of the decomposition input was regarded as a data source, and the associated primary capsule could be marked as corresponding TF or PPI subnetwork. Although the number of the primary capsules was one order of magnitude more than that of previous CapsNet model, the model performed well, and its classification accuracy was comparable with those generated by feed forward neural network and random forest. After training, the contributions of each primary capsule and its corresponding data source to the cell type recognition were revealed by the MultiCapsNet model as co-product of classification. The TF or the PPI subnetwork that labeled the top ranked contributors were often relevant to the cell type they contributed. The comparison of our MultiCapsNet model with SCENIC showed several cell type relevant TFs identified by both methods, which further proves the validity and interpretability of the MultiCapsNet model.

To sum up, our MultiCapsNet model could integrate multiple input sources and standardize the inputs, then use the standardized information for classification through capsule network. In the variant call classification example, the data types are limited to one-hot encoding vectors or real valued vectors. With appropriate dataset, the MultiCapsNet could integrate and standardize more data types, such as sequence data, which can be integrated through convolutional neural network. In addition, our MultiCapsNet model could also incorporate the prior knowledge through adjusting the connection between layers according to the specification of the prior knowledge. In the example of scRNA-seq, we include only PPI and PDI information. In the future, the complex and hierarchical information of biological network will be introduced into the MultiCapsNet model to better understand the intricacies of disease biology (Camacho et al., 2018). Compared with other interpretable machine learning methods, MultiCapsNet could obtain similar classification accuracy under the condition of modular inputs, making it more suitable for the modular biological data.

MultiCapsNet model provides a framework for data integration, especially for multi-omics datasets, which have data from different sources and with different types and formats, or require prior knowledge. Once the data could be transformed into real valued vectors through trainable parameters, the data and transformation process could be integrated into the MultiCapsNet model as a building block. In this sense, the MultiCapsNet model possesses enormous flexibility, and is applicable in many scenes, let alone that it can measure the importance of data sources accompanying the training step without any extra calculation step.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/wanglf19/MultiCapsNet.

AUTHOR CONTRIBUTIONS

JC, JZ, and LW envisioned the project. LW implemented the model and performed the analysis. LW and JC wrote the paper. XM, RN, ZZ, and JZ provided assistance in writing and analysis.

FUNDING

This work was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB38030400 to C.J.); the National Key R&D Program of China (2018YFC0910402 to C.J.); the National Natural Science Foundation of China (32070795 to C.J. and 61673070 to JZ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.767602/full#supplementary-material

REFERENCES

Aebersold, R., and Mann, M. (2016). Mass-spectrometric Exploration of Proteome Structure and Function. *Nature*. 537, 347–355. doi:10.1038/nature19949

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nat. Methods*. 14, 1083–1086. doi:10.1038/nmeth.4463

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., et al. (2018). A Deep Learning Approach to Automate Refinement of Somatic Variant Calling from Cancer Sequencing Data. *Nat. Genet.* 50, 1735–1743. doi:10.1038/s41588-018-0257-y

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning. *Nat. Biotechnol.* 33, 831–838. doi:10.1038/nbt.3300

Jabeen, A., Ahmad, N., and Raza, K. (2018). "Machine Learning-Based State-of-the-Art Methods for the Classification of RNA-Seq Data," *Classification in BioApps. Lecture Notes in Computational Vision and Biomechanicsin*. Editors N. Dey, A. Ashour, and S. Borra (Cham: Springer), vol. 26. doi:10.1007/978-3-319-65981-7_6

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: Accurate Prediction of Single-Cell DNA Methylation States Using Deep Learning. *Genome Biol.* 18, 67. doi:10.1186/s13059-017-1189-z

Benadiba, C., Magnani, D., Niquille, M., Morlé, L., Valloton, D., Nawabi, H., et al. (2012). The Ciliogenic Transcription Factor RFX3 Regulates Early Midline Distribution of Guidepost Neurons Required for Corpus Callosum Development. *Plos Genet.* 8, e1002606. doi:10.1371/journal.pgen.1002606

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. *Cell.* 173, 1581–1592. doi:10.1016/j.cell.2018.05.015

Chen, F., Moran, J. T., Zhang, Y., Ates, K. M., Yu, D., Schrader, L. A., et al. (2016). The Transcription Factor NeuroD2 Coordinates Synaptic Innervation and Cell Intrinsic Properties to Control Excitability of Cortical Pyramidal Neurons. *J. Physiol.* 594, 3729–3744. doi:10.1113/jp271953

Chen, H.-I. H., Chiu, Y.-C., Zhang, T., Zhang, S., Huang, Y., and Chen, Y. (2018). GSAE: an Autoencoder with Embedded Gene-Set Nodes for Genomics Functional Characterization. *BMC Syst. Biol.* 12, 142. doi:10.1186/s12918-018-0642-2

Consortium, E. P. (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*. 306, 636–640. doi:10.1126/science.1105136

Copray, S., Huynh, J. L., Sher, F., Casaccia-Bonnefil, P., and Boddeke, E. (2009). Epigenetic Mechanisms Facilitating Oligodendrocyte Development, Maturation, and Aging. *Glia*. 57, 1579–1587. doi:10.1002/glia.20881

Dan Rosa de Jesus, J. C., Wilson, Rivera., and Crivelli, Silvia. (2018). *Capsule Networks for Protein Structure Classification and Prediction*. arXiv:180807475.

Deng, L., Li, G., Rao, B., and Li, H. (2015). Central Nervous System-specific Knockout of Brg1 Causes Growth Retardation and Neuronal Degeneration. *Brain Res.* 1622, 186–195. doi:10.1016/j.brainres.2015.06.027

Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable Dimensionality Reduction of Single Cell Transcriptome Data with Deep Generative Models. *Nat. Commun.* 9, 2002. doi:10.1038/s41467-018-04368-5

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep Learning: New Computational Modelling Techniques for Genomics. *Nat. Rev. Genet.* 20, 389–403. doi:10.1038/s41576-019-0122-6

Fang, C., Shang, Y., and Xu, D. (2018). Improving Protein Gamma-Turn Prediction Using Inception Capsule Networks. *Sci. Rep.* 8, 15741. doi:10.1038/s41598-018-34114-2

Favot, L., Hall, S. M., Haworth, S. G., and Kemp, P. R. (2005). Cytoplasmic YY1 Is Associated with Increased Smooth Muscle-specific Gene Expression. *Am. J. Pathol.* 167, 1497–1509. doi:10.1016/s0002-9440(10)61236-9

Fukumoto, K., Tamada, K., Toya, T., Nishino, T., Yanagawa, Y., and Takumi, T. (2018). Identification of Genes Regulating GABAergic Interneuron Maturation. *Neurosci. Res.* 134, 18–29. doi:10.1016/j.neures.2017.11.010

Gutiérrez, L., Nikolic, T., van Dijk, T. B., Hammad, H., Vos, N., Willart, M., et al. (2007). Gata1 Regulates Dendritic-Cell Development and Survival. *Blood.* 110, 1933–1941. doi:10.1182/blood-2006-09-048322

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*. 173, 1307. doi:10.1016/j.cell.2018.05.012

Ionescu, A., Kozhemyakina, E., Nicolae, C., Kaestner, K. H., Olsen, B. R., and Lassar, A. B. (2012). FoxA Family Members Are Crucial Regulators of the Hypertrophic Chondrocyte Differentiation Program. *Dev. Cel.* 22, 927–939. doi:10.1016/j.devcel.2012.03.011

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human Protein Reference Database--2009 Update. *Nucleic Acids Res.* 37, D767–D772. doi:10.1093/nar/gkn892

Li, N., Zhao, C. T., Wang, Y., and Yuan, X. B. (2010). The Transcription Factor Cux1 Regulates Dendritic Morphology of Cortical Pyramidal Neurons. *PLoS One* 5, e10596. doi:10.1371/journal.pone.0010596

Li, S., Czubryt, M. P., McAnally, J., Bassel-Duby, R., Richardson, J. A., Wiebel, F. F., et al. (2005). Requirement for Serum Response Factor for Skeletal Muscle Growth and Maturation Revealed by Tissue-specific Gene Deletion in Mice. *Proc. Natl. Acad. Sci.* 102, 1082–1087. doi:10.1073/pnas.0409103102

Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using Neural Networks for Reducing the Dimensions of Single-Cell RNA-Seq Data. *Nucleic Acids Res.* 45, e156. doi:10.1093/nar/gkx681

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep Generative Modeling for Single-Cell Transcriptomics. *Nat. Methods* 15, 1053–1058. doi:10.1038/s41592-018-0229-2

Matsui, M., Sharma, K. C., Cooke, C., Wakimoto, B. T., Rasool, M., Hayworth, M., et al. (2011). Nuclear Structure and Chromosome Segregation in Drosophila Male Meiosis Depend on the Ubiquitin Ligase dTopors. *Genetics* 189, 779–793. doi:10.1534/genetics.111.133819

Molnar, C. (2019). *Interpretable Machine Learning*.

Oakes, S. R., Hilton, H. N., and Ormandy, C. J. (2006). The Alveolar Switch: Coordinating the Proliferative Cues and Cell Fate Decisions that Drive the Formation of Lobuloalveoli from Ductal Epithelium. *Breast Cancer Res.* 8, 207–210. doi:10.1186/bcr1411

Ohata, M., Yamauchi, M., Takeda, K., Toda, G., Kamimura, S., Motomura, K., et al. (2000). RAR and RXR Expression by Kupffer Cells. *Exp. Mol. Pathol.* 68, 13–20. doi:10.1006/exmp.1999.2284

Onodera, K., Fujiwara, T., Onishi, Y., Itoh-Nakadai, A., Okitsu, Y., Fukuhara, N., et al. (2016). GATA2 Regulates Dendritic Cell Differentiation. *Blood J. Am. Soc. Hematol.* 128, 508–518. doi:10.1182/blood-2016-02-698118

Remnestål, J. (2015). *Expression and Distribution of Transcription Factors NPAS3 Och RFX3 in Alzheimer's Disease*. KTH. Skolan för bioteknologi (BIO).

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic Routing between Capsules. *Adv. Neural Inf. Process. Syst.* 30 (Nips), 30. doi:10.1097/01.asw.0000521116.18779.7c

Schulz, M. H., Devanny, W. E., Gitter, A., Zhong, S., Ernst, J., and Bar-Joseph, Z. (2012). DREM 2.0: Improved Reconstruction of Dynamic Regulatory Networks from Time-Series Expression Data. *BMC Syst. Biol.* 6, 104. doi:10.1186/1752-0509-6-104

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., et al. (2014). Single-cell Genome-wide Bisulfite Sequencing for Assessing Epigenetic Heterogeneity. *Nat. Methods.* 11, 817–820. doi:10.1038/nmeth.3035

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a General Repository for Interaction Datasets. *Nucleic Acids Res.* 34, D535–D539. doi:10.1093/nar/gkj109

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an Immeasurable Source of Knowledge. *Contemp. Oncol. (Pozn).* 19, A68–A77. doi:10.5114/wo.2014.47136

Ulven, S. M., Natarajan, V., Holven, K. B., Løvdal, T., Berg, T., and Blomhoff, R. (1998). Expression of Retinoic Acid Receptor and Retinoid X Receptor Subtypes in Rat Liver Cells: Implications for Retinoid Signalling in Parenchymal, Endothelial, Kupffer and Stellate Cells. *Eur. J. Cel. Biol.* 77, 111–116. doi:10.1016/s0171-9335(98)80078-2

Wang, L., Nie, R., Yu, Z., Xin, R., Zheng, C., Zhang, Z., et al. (2020). An Interpretable Deep-Learning Architecture of Capsule Networks for Identifying Cell-type Gene Expression Programs from Single-Cell RNA-Sequencing Data. *Nat. Mach Intell.* 2, 693–703. doi:10.1038/s42256-020-00244-4

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Cell Types in the Mouse Cortex and hippocampus Revealed by Single-Cell RNA-Seq. *Science* 347, 1138–1142. doi:10.1126/science.aaa1934

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. (2018). Deep Learning Sequence-Based Ab Initio Prediction of Variant Effects on Expression and Disease Risk. *Nat. Genet.* 50, 1171–1179. doi:10.1038/s41588-018-0160-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

CORRECTION
published: 27 January 2022
doi: 10.3389/fgene.2021.822045

Corrigendum: MultiCapsNet: A General Framework for Data Integration and Interpretable Classification

Lifei Wang[1,2,3,4], *Xuexia Miao*[2,3], *Rui Nie*[2,3,4], *Zhang Zhang*[5], *Jiang Zhang*[5]* *and Jun Cai*[2,3,4]*

[1]Shulan (Hangzhou) Hospital Affiliated to Zhejiang Shuren University Shulan International Medical College, Hangzhou, China, [2]China National Center for Bioinformation, Beijing, China, [3]Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China, [4]University of Chinese Academy of Sciences, Beijing, China, [5]School of Systems Science, Beijing Normal University, Beijing, China

Keywords: capsule network, classification, data integration, interpretability, modular feature

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

***Correspondence:**
Jiang Zhang
zhangjiang@bnu.edu.cn
Jun Cai
juncai@big.ac.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *25 November 2021*
Accepted: *26 November 2021*
Published: *27 January 2022*

Citation:
Wang L, Miao X, Nie R, Zhang Z, Zhang J and Cai J (2022) Corrigendum: MultiCapsNet: A General Framework for Data Integration and Interpretable Classification. Front. Genet. 12:822045. doi: 10.3389/fgene.2021.822045

A corrigendum on

MultiCapsNet: A General Framework for Data Integration and Interpretable Classification
by Wang, L., Miao, X., Nie, R., Zhang, Z., Zhang, J., and Cai, J. (2021). Front. Genet. 12:767602. doi:10.3389/fgene.2021.767602

In the original article, there was a mistake in the number labeling for **Figure 3** and **Figure 4** as published. **Figure 3** should be labeled as **Figure 4**, and vice versa. The correct legend appears below.

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

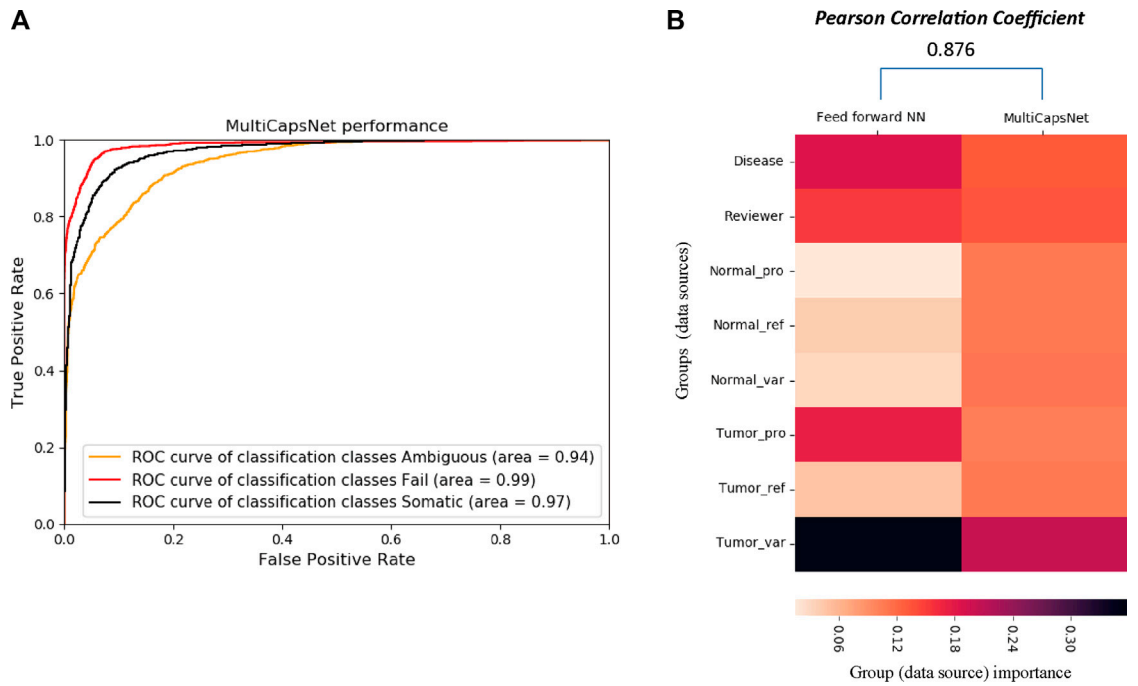


FIGURE 3 | Architecture of MultiCapsNet integrated with prior knowledge. **(A)** The model has two layers. The first layer consists of 696 parallel neural networks corresponding to 696 primary capsules labeled with either transcription factor (348) or protein-protein interaction cluster node (348). The inputs of each primary capsule include genes regulated by a transcription factor or in a protein-protein interactions sub-network. The second layer is the Keras implementation of CapsNet for classification. The length of each final layer type capsule represents the probability of input data belonging to the corresponding classification category. **(B)** Alternative representation of MultiCapsNet integrated with prior knowledge. Genes that are regulated by a transcription factor or in a protein-protein interactions sub-network, are groups together as a data source for MultiCapsNet. **Figures 3A,B** are equivalent with different representation.

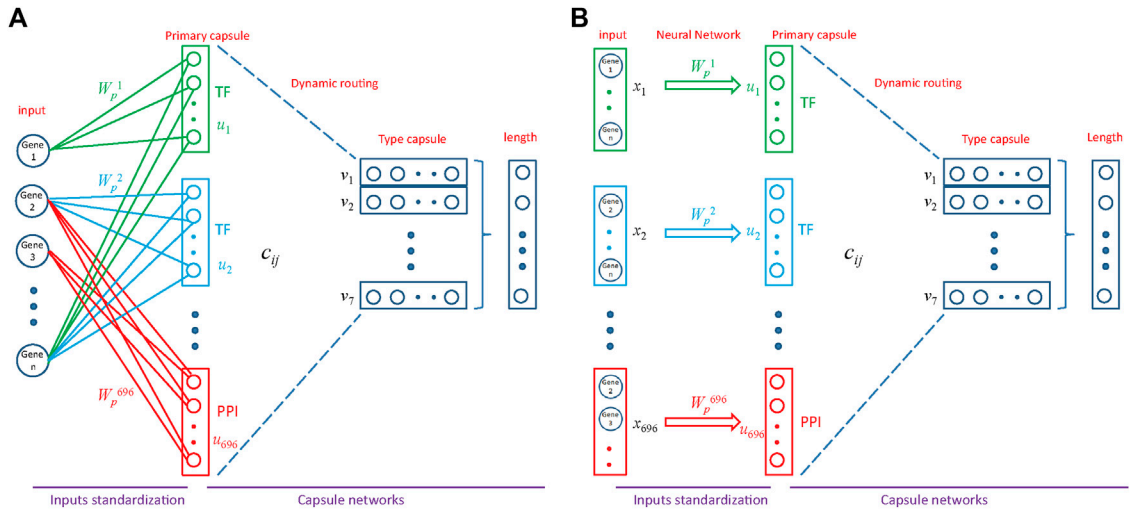


FIGURE 4 | The comparison between MultiCapsNet and feed forward neural network shows the high performance and interpretability of MultiCapsNet. **(A)** The AUC scores demonstrate that the MultiCapsNet model achieves very high classification performances in all three classification categories. **(B)** The normalized group (data source) importance scores generated by MultiCapsNet and feed forward neural network are highly correlated.

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 February 2022
doi: 10.3389/fgene.2021.806457

Machine Learning of Single Cell Transcriptomic Data From anti-PD-1 Responders and Non-responders Reveals Distinct Resistance Mechanisms in Skin Cancers and PDAC

Ryan Liu[1], *Emmanuel Dollinger*[1,2,3,4]* *and Qing Nie*[1,2,3,4]*

[1]Department of Mathematics, University of California, Irvine, Irvine, CA, United States, [2]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA, United States, [3]Center for Complex Biological Systems, University of California, Irvine, Irvine, CA, United States, [4]NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, Irvine, CA, United States

OPEN ACCESS

Edited by:
Xiaofei Zhang,
Central China Normal University, China

Reviewed by:
Jinzhi Lei,
Tianjin Polytechnic University, China
Xiaoqi Zheng,
Shanghai Normal University, China

***Correspondence:**
Emmanuel Dollinger
emmanuel.dollinger@uci.edu
Qing Nie
qnie@uci.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 31 October 2021
Accepted: 16 December 2021
Published: 01 February 2022

Citation:
Liu R, Dollinger E and Nie Q (2022) Machine Learning of Single Cell Transcriptomic Data From anti-PD-1 Responders and Non-responders Reveals Distinct Resistance Mechanisms in Skin Cancers and PDAC. Front. Genet. 12:806457. doi: 10.3389/fgene.2021.806457

Immune checkpoint therapies such as PD-1 blockade have vastly improved the treatment of numerous cancers, including basal cell carcinoma (BCC). However, patients afflicted with pancreatic ductal carcinoma (PDAC), one of the deadliest malignancies, overwhelmingly exhibit negative responses to checkpoint therapy. We sought to combine data analysis and machine learning to differentiate the putative mechanisms of BCC and PDAC non-response. We discover that increased MHC-I expression in malignant cells and suppression of MHC and PD-1/PD-L expression in $CD8^+$ T cells is associated with nonresponse to treatment. Furthermore, we leverage machine learning to predict response to PD-1 blockade on a cellular level. We confirm divergent resistance mechanisms between BCC, PDAC, and melanoma and highlight the potential for rapid and affordable testing of gene expression in BCC patients to accurately predict response to checkpoint therapies. Our findings present an optimistic outlook for the use of quantitative cross-cancer analyses in characterizing immune responses and predicting immunotherapy outcomes.

Keywords: immunotherapy, machine learning of single cell sequencing, therapeutic response prediction, supervised learning, deep learning, single-cell transcriptomic sequencing, basal cell carcinoma, pancreatic ductal adenocarcinoma

1 INTRODUCTION

Cancer immunotherapy has shown to elicit substantial response to many cancers and has led to significant increases in quality of life for cancer patients. This is especially true of checkpoint therapy, which causes tumor regression in previously untreatable cancers. Response to checkpoint therapy has been positively correlated with tumor mutational burden (TMB) and with presence of $CD8^+$ T cells in the tumor microenvironment (characterized as "hot" tumors) (Yarchoan et al., 2017) (Tumeh et al., 2014). However, the potential mechanisms of checkpoint therapy are still being investigated and there are as of yet few prognostic markers for response (Bai et al., 2020). Potential

biomarkers include the alteration of signaling pathways in tumor cells, namely mutations in the interferon (IFN)-γ pathway, as well as pathways related to tumor cell proliferation and infiltration (Possick, 2017). Poor response to immunotherapy is also linked to inactivation of *PTEN*, mutations of *POLE*, and linked mutations in *KRAS* and *STK11* (Wang et al., 2021).

Basal cell carcinoma (BCC) is a skin cancer with high TMB (estimates range from a median of 47.3 mutations/Mb to 65 mutations/Mb), arising from skin membrane stem cells (Chalmers et al., 2017) (Bonilla et al., 2016). Despite being mostly characterized as a "cold" tumor, BCC has been shown to exhibit partial and complete responses to checkpoint therapy (Moujaess et al., 2021) (Walter et al., 2010). Recently, the PD-1 inhibitor cemiplimab received FDA approval for patients with advanced non-resectable BCC that are resistant to Hedgehog pathway inhibition (Moujaess et al., 2021). BCC is a relatively unique cancer in that the TMB does not correlated with immunogenicity. This is thought to be a combination of downregulation of major histocompatability complex class I (MHC-I) expression and immunosuppression via influx of T regulatory cells driven by overexpression of the Hedgehog pathway (Walter et al., 2010) (Grund-Gröschke et al., 2020).

Pancreatic ductal adenocarcinoma, a carcinoma arising from ductal cells in the pancreas, is in essence an incurable disease, with less than 5% survival rate over 5 years as of 2016 (Bengtsson et al., 2020). This survival rate, in conjunction with projections that pancreatic cancers will be one of the major causes of cancer-related deaths by 2030, highlights a strong need to develop better biomarkers and treatments (Rahib et al., 2014). Despite significant progress having been made in oncology treatment, PDAC has proven to be incredibly challenging to treat, due to a multitude of factors including lack of symptoms before metastasis and lack of specific clinical characteristics (Wolfgang et al., 2013). PDAC has also been found to be non-responsive to checkpoint immunotherapy, showing a poor response to CTLA-4, PD-1 and PD-L1 therapies (Royal et al., 2010) (Renouf et al., 2020). The reasons for this lack of response are still under study; proposed factors include levels of microsatellite instability, tumor infiltrating lymphocytes (TILs), and DNA mismatch repair deficiency (Christenson et al., 2020) (Pu et al., 2019). Although it has a relatively low TMB, PDAC has a highly immunosuppressive tumor microenvironment and is immunogenic (Fan et al., 2020).

In order to study the differential mechanisms by which BCC and PDAC cancers resist checkpoint immunotherapy treatment and building on our previous work (Dollinger et al., 2020), we leveraged two recent single-cell transcriptomic datasets of PDAC and BCC (**Figures 1A** and **Supplementary Figure S1**). Through comparing these two datasets, we identified potential common biomarkers for nonresponse to PD-1 blockade and differences in the immune mechanisms combating tumor progression in these two cancers. We found that PDAC suppresses MHC-I gene expression in $CD8^+$ T cells and upregulates MHC-I in malignant cells compared to BCC. Furthermore, the PD-1/PD-L signaling axis is significantly weaker in PDAC, leaving diminished opportunity for phenotypic changes to occur through boosting its activity. Utilizing machine learning classification algorithms, we additionally discovered that PDAC displays greater similarities to melanoma, which is highly immunogenic and undergoes rapid metastasis, than to BCC (Dollinger et al., 2020).

2 RESULTS

2.1 Characterization of the BCC and PDAC TME

In order to characterize the transcriptomic differences between responders and non-responders to PD-1 blockade therapy, we analyzed a previously published scRNA-seq dataset of basal cell carcinoma patients pre- and post-treatment (Yost et al., 2019). The dataset consists of 24 site-matched samples from 11 patients with advanced BCC; a total of 53,030 malignant, immune, and stromal cells were obtained between the six responsive and five nonresponsive patients. Unsupervised clustering of the dataset revealed 20 distinct clusters (**Figure 1B**), including 8 T cell clusters and two malignant cell clusters (Methods). Our clustering largely agrees with the original analysis (**Supplementary Figure S2A**), with the exceptions that we only found 1 B cell cluster and differentiated macrophages into the M1/M2 polarization as defined in (Orecchioni et al., 2019).

Separately, a dataset of 46,244 cells from 16 PDAC patients and 8,541 cells from three non-malignant adjacent samples was used to characterize the PDAC TME (Steele et al., 2020); all samples were taken before any treatment and include both surgical and fine-needle biopsy specimens. Both the malignant and adjacent samples were integrated together before clustering, which revealed 22 distinct subpopulations (**Figure 1C**). Whereas the general cluster labels correspond with those of the original paper, two important distinctions are made. First, $CD8^+$ T cells are divided into effector/activated cells, memory cells, and chronically activated/exhausted cells, referred to hereafter as exhausted cells; these labels correspond with the $CD8^+$ T cell subclusters in the BCC dataset to facilitate further direct comparison, and are therefore not equivalent to those in Extended Data **Figure 4** of the original analysis. However, examination of mean scaled expression of highly enriched genes reveals that the newly defined clusters are transcriptomically similar to those in the original analysis (**Supplementary Figure S2B**). Second, within the population of epithelial/ductal cells, two distinct clusters of malignant cells were identified using 205 marker genes commonly upregulated in PDAC tumor samples (**Figure 1D**) (Tang et al., 2018). The identification of these clusters is novel and was not detected by the original authors. Whereas one malignant cluster had significantly elevated expression of nearly all marker genes and a high percentage (>50%) of all cells expressing each gene, the second malignant cluster had much more sparse and less significantly elevated expression of the DEGs, suggesting that there exists a wide spectrum in the degree of malignancy of ductal cells (**Supplementary Figure S2C**). Both normal ductal cells in

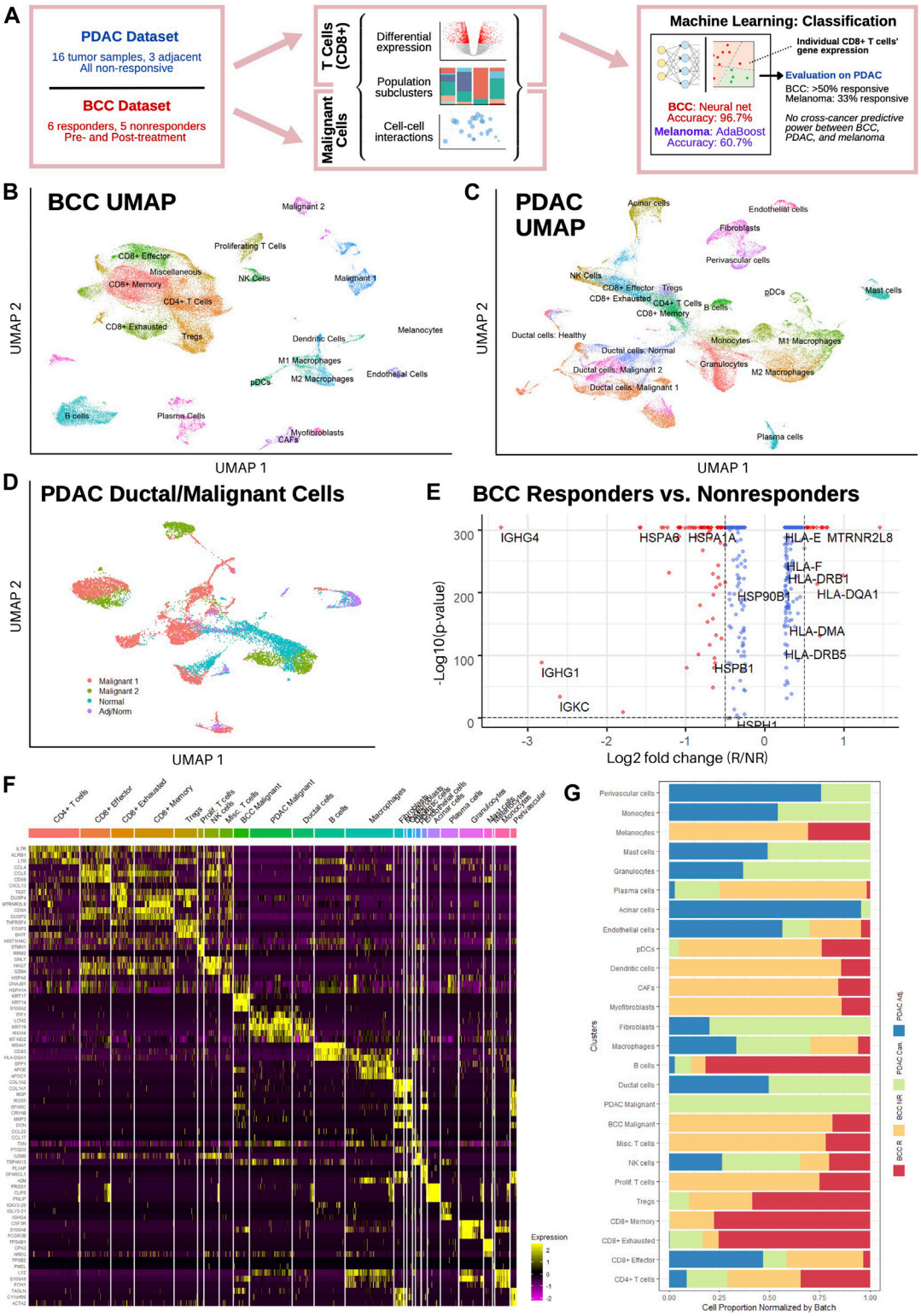


FIGURE 1 | Single-cell sequencing reveals distinct T cell subpopulations in BCC and PDAC and ductal cell subpopulations in PDAC. **(A)** Workflow diagram. **(B,C)** Dimensional reduction of **(B)** BCC and **(C)** PDAC TME. **(D)** Dimensional reduction of PDAC ductal cells. **(E)** Differential gene expression between BCC responders and nonresponders; positive fold change indicates greater expression in responders. **(F)** Single-cell resolution heatmap of top three differentially expressed genes per cluster in merged BCC and PDAC dataset. **(G)** Normalized proportion of cells in each cluster identified in **(F)** that belong to BCC responders (BCC R), BCC nonresponders (BCC NR), PDAC tumors (PDAC Can.), and adjacent PDAC samples (PDAC Adj.).

malignant PDAC samples and those from adjacent samples exhibited negligible expression of the 205 marker genes.

Comparing average gene expression across all cells between responders and nonresponders in BCC, we found that MHC genes are overexpressed in responders, whereas heat shock protein (*HSP*) genes are overexpressed in nonresponders (**Figure 1E**). This is in line with current literature: reduced MHC-I expression is well-known to facilitate immune evasion (Šmahel, 2017); MHC-II expression is correlated with response to PD-1 blockade treatment (Rodig et al., 2018); and *HSP* genes are associated with tumor proliferation and metastasis (Ciocca and Calderwood, 2005). Merging both datasets, we find that the top differentially expressed genes (DEGs) for each cluster aligns with the marker genes used to identify them in (Yost et al., 2019) and (Steele et al., 2020) (**Figure 1F**, Methods). Furthermore, no significant difference was detected in the expression of top DEGs in each cluster, e.g. expression of *CCL4*, *CCL5*, and *CD59* is similar between PDAC and BCC $CD8^+$ effector T cells. However, wide discrepancies can be seen in the relative populations of different clusters between BCC responders, BCC nonresponders, and PDAC patients (**Figure 1G**). We find that BCC responders are more heavily represented amongst B cells and T cells, whereas BCC nonresponders have greater numbers of stromal, myeloid, and malignant cells, recapitulating previous analyses (Dollinger et al., 2020) (Yost et al., 2019). Meanwhile, PDAC tumors have very low numbers of B cells and T cells in comparison to all BCC tumors, but have much larger populations of macrophages and endothelial cells. This highlights the challenges of using immunotherapy in PDAC; it also justifies the comparison of two different cancers due to the similarities in cell population between non-responders in BCC and PDAC.

2.2 CD8+ T Cells Are More Active in BCC Than PDAC

One of the main functions of PD-1 blockade is to reinvigorate exhausted $CD8^+$ T cells, leading to a stronger anti-tumor response and eventual tumor regression (Verma et al., 2019); thus, the altered function and composition of T cells is a primary suspect in the nonresponse of PDAC to immunotherapies. Due to the absence of data on PDAC responders, in this section we compare T cells in PDAC to those in BCC responders and nonresponders. Similarities in composition or gene expression between the T cells of PDAC and BCC nonresponders, as well as commonalities in the differences between BCC responders and nonresponders and the differences between BCC responders and PDAC, provide potential factors for further study.

Comparing the T cell populations in PDAC tumor sites and adjacent samples, we find significant differences in relative subpopulation sizes – in particular, there are virtually no regulatory T cells (Tregs), memory $CD8^+$ T cells, or exhausted $CD8^+$ T cells in adjacent samples (**Figure 2A**). This suggests that a subset of effector $CD8^+$ T cells in the PDAC TME enter an exhausted phenotype or differentiate into memory cells after prolonged exposure to cancer (Xia et al., 2019). Furthermore, we unexpectedly find a substantially larger population of $CD8^+$ exhausted and memory cells in BCC responders and a diminished number of $CD8^+$ effector cells. Therefore, it is possible that BCC responders benefit more from PD-1 therapy due to greater potential for phenotypic shifts on the cellular level, or simply that T cells in responders have experienced prolonged exposure to the malignancy. Both PDAC populations lack proliferating T cells, supporting its reputation as extremely immunosuppressive (Foucher et al., 2018).

To determine whether these trends are patient-specific, we first compared the fraction of all cells in each pre-treatment sample that are identified as T cells (**Figure 2B**). As expected, pre-treatment responders have a greater proportion of T cells than nonresponders, although the difference is statistically insignificant ($p > 0.05$). However, both BCC responders and nonresponders have significantly higher T cell proportions than both malignant and adjacent PDAC samples by a factor of 4–8. Comparing the proportion of T cells classified as activated and exhausted, we find that the proportions are similar between all patients, with the unexpected exception that the vast majority of T cells in the adjacent pancreas samples are effector $CD8^+$ T cells (**Figures 2C, D**). This may indicate that adjacent samples may not reflect a true negative control, as is often used in the literature.

Identification of the top 2,000 genes with the most similar gene expression between BCC and PDAC unsurprisingly reveal no notable gene groups, supporting the theory that the two cancers rely on different systems of immune activation. However, we notice that *HLA* genes are amongst the most highly enriched genes in both cancers; furthermore, they are consistently overexpressed in BCC compared to PDAC by a factor of 2-10 with the exception of *HLA-E* and *HLA-F*, suggesting that PDAC suffers from much more severe MHC-I suppression (**Figure 2E**). To confirm whether these differences hold on a patient level, we constructed a MHC-I and MHC-II score (Methods). Comparison of the per-patient MHC-I scores between BCC and PDAC for each of the $CD8^+$ T cell subclusters shows that regardless of the subcluster, BCC $CD8^+$ T cells have significantly elevated MHC-I expression in comparison to PDAC; this discrepancy is most pronounced in memory $CD8^+$ T cells, where MHC-I scores are on average 4 times higher (**Figure 2F**). Similarly, per-patient comparison of MHC-II scores show that all three groups of $CD8^+$ T cells have significantly lower expression in PDAC than BCC - in particular, the majority of effector and memory $CD8^+$ T cells in PDAC exhibit virtually no MHC-II expression (**Figure 2G**). This supports prior research demonstrating that MHC-I molecules are degraded by autophagy-dependent mechanisms in PDAC, thereby facilitating impaired antigen presentation and resistance to checkpoint therapies (Johnson et al., 2016); no such mechanisms have been implicated in BCC and this provides evidence against such a mechanism existing in the BCC TME.

We then compared the distribution of PD-1 and PD-L1/PD-L2 expression in BCC and PDAC $CD8^+$ T cells (**Figures 2H, I**). In both cancers, the vast majority (>95%) of cells exhibit zero

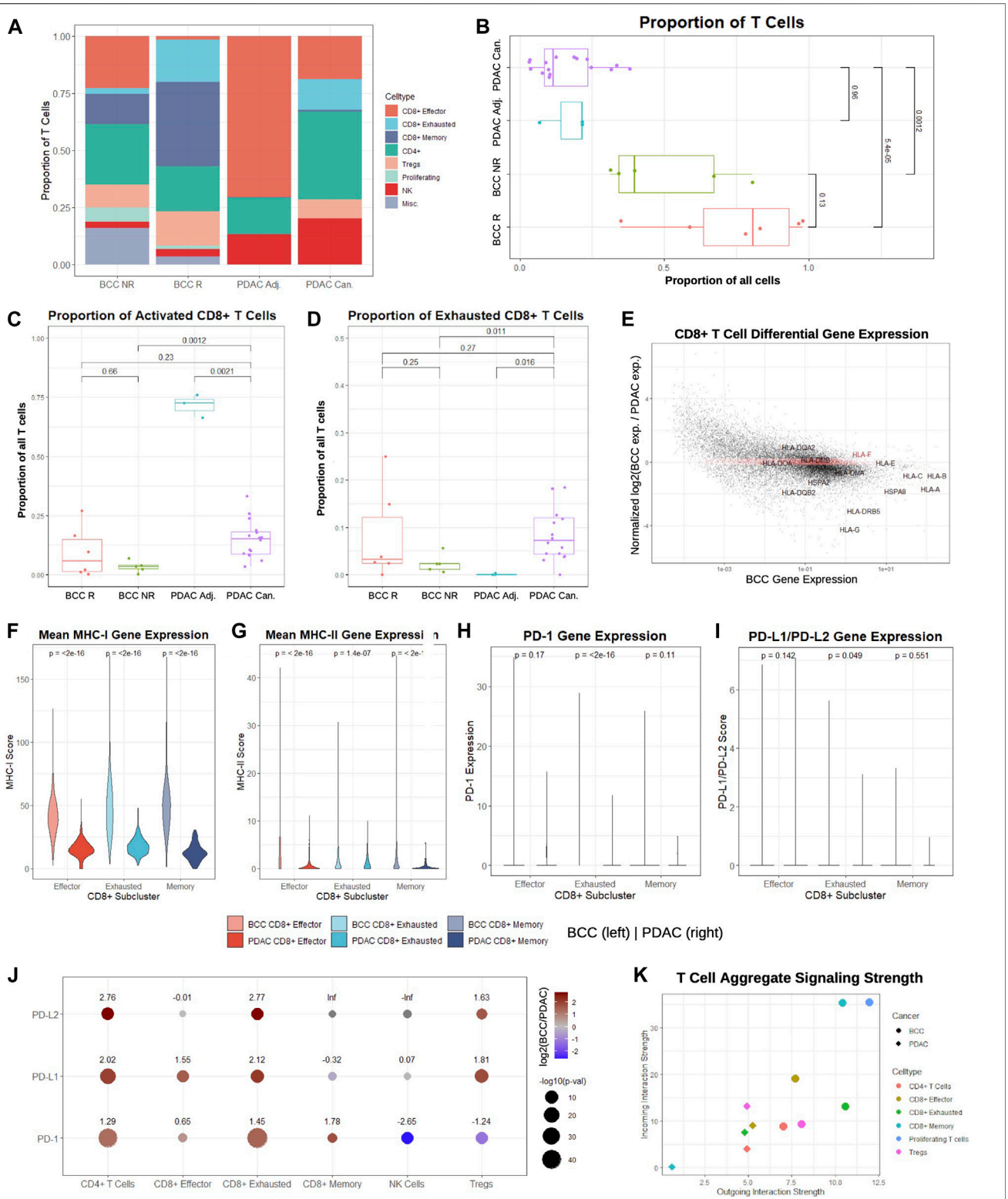


FIGURE 2 | CD8$^+$ T cells in BCC exhibit higher signaling strength and MHC-I/PD-1 pathway expression compared to PDAC. **(A)** Proportion of T cell subclusters in BCC (R = Responsive, NR = Nonresponsive) and PDAC (Can. = Cancerous, Adj. = Adjacent) samples. **(B)** Per-patient comparison of the proportion of all cells that are T cells. **(C, D)** Per-patient comparison of the proportion of T cells that are **(C)** CD8$^+$ effector and **(D)** CD8$^+$ exhausted. **(E)** Differential gene expression in CD8$^+$ T cells, BCC vs PDAC (positive fold change indicates higher expression in BCC); the top 2000 genes with the lowest absolute value normalized fold change in expression are highlighted in red. **(F–I)** Per-cell comparison between BCC and PDAC CD8$^+$ T cell subclusters on **(F)** mean MHC-I, **(G)** mean MHC-II, **(H)** PD-1, and **(I)** mean PD-L1/PD-L2 gene expression. **(J)** Comparison of fold change between average PD-1 pathway gene expression of BCC and PDAC T cell subclusters; positive values indicate greater expression in BCC. **(K)** Comparison of aggregate outgoing and incoming signaling strength between BCC and PDAC T cell subclusters.

expression in all subclusters; the only exception is exhausted $CD8^+$ T cells in BCC. Thus, no significant difference ($p > 0.01$) is detected between BCC and PDAC in the expression distribution of the PD-1 pathway. Due to the exceptionally low expression in all clusters, we examined the mean expression of all cells in each T cell subcluster (**Figure 2J**). We find that with the exception of NK cells and Tregs in the expression of PD-1, BCC T cells are also upregulated in PD-1/PD-L1/PD-L2 in all subclusters by a factor of 3–7. These results imply that a combination of immune suppression and low expression of the MHC-I and PD-1 gene pathway in $CD8^+$ T cells contribute to decreased response to PD-1 blockade in PDAC, as there is not sufficient expression of the PD-1 pathway to induce significant change in T cell activity with PD-1 blockade.

Lastly, we compared the aggregate signaling strength of each BCC and PDAC T cell subcluster to determine their level of intercellular communication (**Figure 2K**, Methods). Signaling strength was inferred using the CellChat package by considering multiple measures of network centrality for each cluster, utilizing a manually-curated database of hundreds of ligand-receptor interactions (Jin et al., 2021). We find that nearly all BCC T cell subclusters are more dominant "senders" and "receivers" than their PDAC counterparts. In particular, due to the small size of the $CD8^+$ memory T cell population in PDAC, it exhibits negligible intercellular communication, whereas $CD8^+$ memory T cells in PDAC are extremely active. Additionally, proliferating T cells in BCC are the dominant senders and receivers, despite constituting 4% of the BCC T cell population; no equivalent subcluster was identified in the PDAC samples. Altogether, these results demonstrate that $CD8^+$ T cells in BCC are substantially more active than their counterparts, both in aggregate and in the MHC and PD-1 pathways.

2.3 Differential Expression of MHC-I in Malignant Cells Is Associated With Response to PD-1 Therapy

Multiple distinct subtypes of PDAC have been defined on the basis of significant inter-tumoral and intra-tumoral heterogeneity to develop personalized treatment strategies (Moffitt et al., 2015). To determine whether the two distinct malignant ductal cell subpopulations in our clustering (see **Figure 1E**) represent unique subtypes, we compared the marker genes for each PDAC ductal cell subcluster (**Figure 3A**). We find that the majority of the top markers for normal ductal cells in cancerous patients are mitochondrial genes (*MT-ND2*, *MT-ND1*, *MT-ND4*, *MT-CO1*, *MT-ND5*, *MT-ATP6*, *MT-CO3*, *MT-CYB*, *MT-ND3*, *MT-CO2*, *MTRNR2L12*, *MT-ND6*, and *MT-ND4L*), supporting previous research that mitochondrial metabolic reprogramming may be crucial to the progression of pancreatic cancers (Reyes-Castellanos et al., 2020). The Malignant 2 cluster was characterized by upregulation of several ribosomal protein (RP) genes, corroborating hypotheses that unique RP transcript expression can be utilized in defining unique cancer subtypes (Dolezal et al., 2018). Interestingly, ductal cells from adjacent samples exhibited elevated expression of marker genes for the Malignant 1 subcluster in comparison to ductal cells from cancerous samples. Although each malignant cell subcluster is dominated by a subset of the cancerous samples (**Supplementary Figure S3A**), suggesting that inter-tumoral heterogeneity led to the presence of two distinct PDAC subtypes in the dataset, we fail to identify concrete evidence linking either malignant cluster to the cancer subtypes defined in (Moffitt et al., 2015) and (Bailey et al., 2016).

While numerous studies have been conducted on the genetic markers of BCC and PDAC individually (Pellegrini et al., 2017) (Liu et al., 2020) (Tang et al., 2018) (Kunovsky et al., 2018), we present here the first direct comparison between the expression patterns of malignant cells in the two cancers (**Figure 3B**). We find that unsurprisingly keratins (namely *KRT5*, *KRT14*, and *KRT17*) which are marker genes for keratinocytes are overexpressed in BCC compared to PDAC by approximately an order of magnitude. Furthermore, many *HSP* genes are unexpectedly upregulated in BCC malignant cells, despite their well-known association with carcinogenesis and metastasis (Ciocca and Calderwood, 2005) (Wu et al., 2017). Expression of *HLA* genes (MHC-I and MHC-II) are slightly upregulated in PDAC malignant cells. The most upregulated genes in PDAC include *SPINK1*, known for contributing towards increased tumor proliferation and poor cancer prognosis (Mehner and Radisky, 2019); *TFF1*, which facilitates PDAC metastasis (Arumugam et al., 2011); and *S100A6*, a key diagnostic marker for PDAC (Leclerc and Vetter, 2015).

As the role of MHC-I and MHC-II in both BCC and PDAC tumors are well-established, we sought to compare the distribution of *HLA* gene expression between the two cancers. Using the same calculation for the MHC-I and MHC-II scores as **Figures 2F–H**, we find that on a cellular level, MHC-I expression is significantly upregulated in PDAC malignant cells compared to normal ductal cells from both cancerous and adjacent samples (**Figure 3C**). Interestingly, ductal cells from adjacent samples also had significantly higher MHC-I expression than those from normal samples. This provides more evidence that adjacent samples are not true negative controls. MHC-II expression followed the same trends, notably with a majority of non-cancerous ductal cells having zero expression (**Figure 3D**).

Looking at average MHC-I and MHC-II expression per patient between malignant and normal ductal cells, similar trends emerge. Whereas MHC-I expression is significantly elevated in malignant cells, no significant difference exists between average MHC-II scores, with the distribution of scores for normal ductal cells actually possessing a higher median and much greater variance (**Supplementary Figure S3C**). This suggests that there exists significant inter-tumoral variability in MHC-II expression, with the larger tumor samples having lower expression and therefore disproportionately shifting the cellular distribution downwards.

It appears that PDAC is non-responsive to treatment despite having already-elevated levels of MHC-I expression. To determine the relationship between MHC-I expression and response, we turned our attention to analyzing differences in MHC expression between BCC responders and nonresponders, both before and after treatment. Surprisingly, we find that

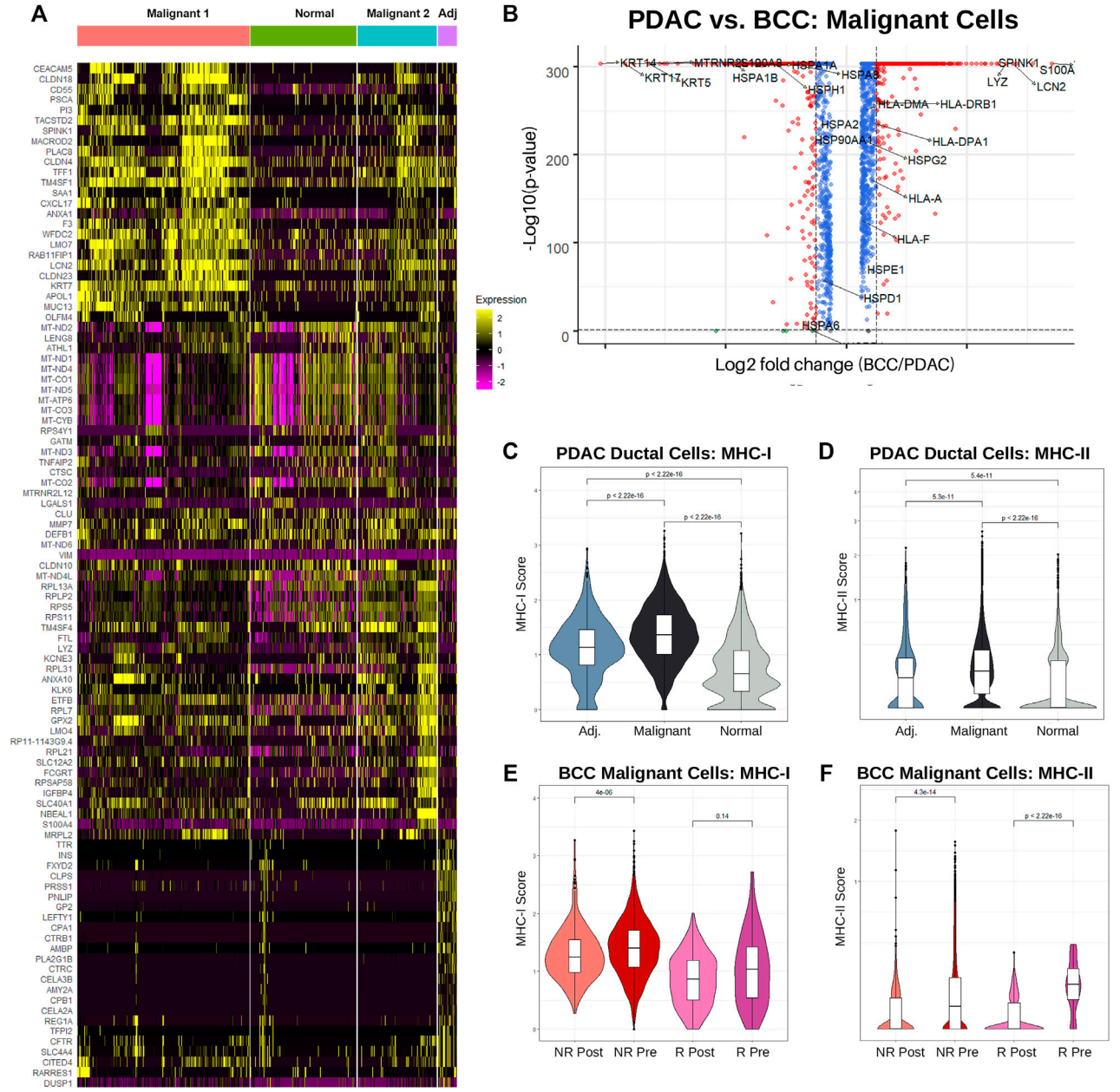


FIGURE 3 | Increased MHC-I expression is associated with nonresponse to PD-1 therapy in BCC and PDAC. **(A)** Single-cell-resolution heatmap of top 20 most differentially expressed genes in each PDAC ductal cell subcluster. **(B)** Differential gene expression between BCC and PDAC malignant cells; higher gene expression in PDAC is denoted by a positive log2 fold change. **(C)** Violin plot of MHC-I score by cell in ductal cells. **(D)** Differential gene expression in malignant cells, PDAC vs BCC. **(E, F)** Violin plot of the **(E)** MHC-I and **(F)** MHC-II scores per cell in BCC malignant cells, classified by response (R/NR) and treatment (Pre/Post).

regardless of response status, MHC-I expression slightly decreased post-treatment; however, nonresponders overall had higher expression (**Figure 3E**). This is likely due to a combination of greater initial tumor malignancy in non-responders and T cell exhaustion over time. Meanwhile, responders had significantly higher MHC-II expression pre-treatment, but both responders and non-responders experienced drastic reductions in expression post-treatment (**Figure 3F**). Whereas PDAC malignant cells exhibited greater similarities in MHC-I expression to BCC non-responders, MHC-II expression was more similar to BCC responders.

2.4 Machine Learning Reveals Divergent Immune Mechanisms in Response to PD-1 Blockade

With stark differences in immunogenicity, TMB, and tumor progression between BCC and PDAC, it is hardly surprising

that the immune mechanisms implicated through PD-1 blockade in the two cancers are completely divergent. However, there exists greater similarity in the immunosuppressivity and immunogenicity between PDAC and melanoma, which exhibits a relatively high response rate to PD-1 blockade of 30–45% (Sun et al., 2020) (Ribas and Wolchok, 2018). To test whether the immune response of PDAC is more similar to BCC or melanoma, and whether differential gene expression can recapitulate these differences, we turn to machine learning. $CD8^+$ cytotoxic T cells are most directly responsible for killing tumor cells and constitute the largest cluster in our datasets. Therefore, in this section we attempt to construct a supervised learning algorithm to predict whether individual $CD8^+$ T cells originate from a patient responsive or nonresponsive to treatment. Patients with a high percentage of $CD8^+$ T cells predicted to be responsive will therefore have a higher likelihood of response to PD-1 blockade.

Separate supervised learning algorithms were trained on both BCC and melanoma $CD8^+$ pretreatment T cells, subsetted from (Yost et al., 2019) and (Sade-Feldman et al., 2018) respectively. The BCC dataset consists of 4,311 cells (229 effector, 1,104 exhausted, and 2,978 memory) from five responders and 1,571 cells (73 effector, 2,439 exhausted, 4,156 memory) from six nonresponders; the melanoma dataset consists of 1,512 cells from 17 responsive samples and 1,239 cells from 31 nonresponsive (32 total patients). The cells from (Sade-Feldman et al., 2018) were FACS sorted on $CD45^+$ before plating and sequencing. Identification of $CD8^+$ T cells in the melanoma dataset were taken directly from (Sade-Feldman et al., 2018) and (Dollinger et al., 2020).

Each dataset was first filtered to include only genes with expression detected in all three datasets (BCC, melanoma, and PDAC). Classifiers were then constructed on the BCC and melanoma $CD8^+$ T cells through the sci-kit learn pipeline (Pedregosa et al., 2011), using only the top 2,000 highly variable genes in each dataset respectively (**Figures 4A, B**, Methods). Through benchmarking multiple classifiers against one another, we are able to identify the classification algorithm which most accurately responds to the features present in our datasets. With the exception of Naive Bayes, all classifiers demonstrated high training accuracy (>73%) on the BCC dataset; the best model was the multilayer perceptron (MLP) neural network, which achieved 96.7% testing accuracy on the original dataset after parameter optimization. Classifiers trained on the melanoma dataset were noticeably weaker, with training accuracy between 50 and 62%; after optimization, the best model was the AdaBoost, which achieved 60.7% testing accuracy. This could stem from the extremely high intratumor and intertumor heterogeneity observed in melanoma, which lowers predictive power (Grzywa et al., 2017). In addition, many of the melanoma patients were previously treated with other chemotherapeutics, potentially altering the immune environment and confounding the classification of responders.

To guard against overfitting, classifiers utilizing a lower number of most highly variable genes were constructed for both BCC and melanoma. Remarkably, in both datasets, predictive power remained notably strong until using just 20 or less genes. In particular, the 20-gene BCC classifier with 81% accuracy utilized *CXCL13*, *HSPA1A*, *HSPA6*, *HSPA1B*, *G0S2*, *XCL1*, *CCL4*, *FOS*, *GNLY*, *TRBV11-2*, *XCL2*, *KRT86*, *NMB*, *DNAJB1*, *CCL4L2*, *SOX4*, *ID3*, *HSP90AA1*, *NR4A1*, and *MT1G* (**Supplementary Figure S4A, B**). This suggests that traditional gene expression tests may be used as a quick predictor of response to PD-1 blockade with reasonably accuracy (60–80%) in BCC and melanoma.

Melanoma and BCC are known to exhibit very different immune mechanisms: whereas melanoma is immunogenic and demonstrates resistance to immunotherapy, BCC is relatively non-immunogenic and suffers from low immune cell recruitment and activation (Dollinger et al., 2020). To confirm these differences on a transcriptomic level, we tested the BCC classifier on the melanoma dataset and vice versa. As expected, both classifiers performed similarly to random chance (AUC = 0.501 and 0.486 respectively), providing support for the different immune evasive and suppressive mechanisms of the cancers' response to PD-1 blockade (**Figure 4C**).

On a per-patient resolution, the vast majority of BCC $CD8^+$ T cells from responders or nonresponders are classified as responsive or nonresponsive by the BCC neural net respectively (**Figure 4D**) – there exists a significant difference in the proportion of cells that are responsive between responders and nonresponders ($p = 0.004\,35$). However, when applied to PDAC cells, slightly over half of the cells were declared responsive. Meanwhile, surprisingly no significant distinction ($p = 0.13$) can be made in comparing the percentage of cells classified as responsive between melanoma responders and nonresponders (**Figure 4E**), with several responders having only a small fraction of cells being classified as such. This indicates that construction of the classifier was likely biased towards samples with larger numbers of $CD8^+$ T cells. The melanoma classifier furthermore identifies a mean of 33% of cells in a PDAC patient as nonresponsive, similar to melanoma nonresponders ($p = 0.46$) and significantly lower than responders ($p = 0.099$), although there exists significantly inter-patient variability. Under the assumption that the vast majority of PDAC patients would not respond to PD-1 blockade, it is evident that the melanoma classifier performs markedly better on the PDAC dataset than the BCC classifier. This suggests that similarities between resistance mechanisms between melanoma and PDAC may extend to $CD8^+$ T cells in addition to macrophages (Zhu et al., 2014).

3 DISCUSSION

To date, multiple studies of BCC have established its relative ease in prognosis and treatment; meanwhile, PDAC continues to evade early-stage detection and exhibits uniformly poor response to existing checkpoint immunotherapies. Consistent with existing literature, our direct comparison of the BCC and PDAC TMEs reveal that PDAC tumors foster a more immunosuppressive microenvironment compared to BCC (Foucher et al., 2018). In particular, although BCC is known to downregulate MHC-I expression (Dhatchinamoorthy et al.,

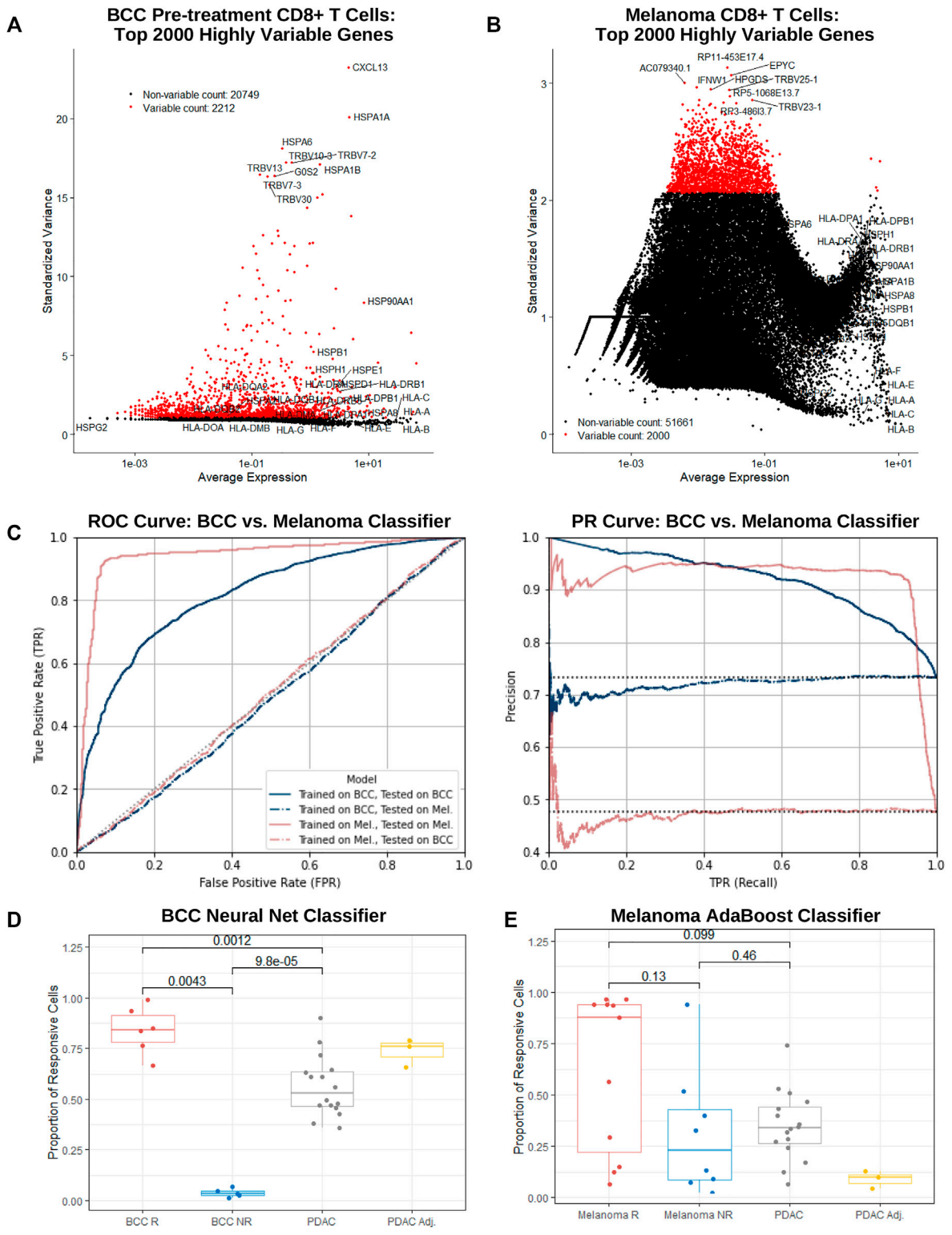


FIGURE 4 | BCC, PDAC, and melanoma exhibit different immune mechanisms in response to PD-1 blockade. **(A,B)** Dot plot of average gene expression vs standardized variance of all CD8$^+$ T cells in **(A)** BCC pre-treatment and **(B)** melanoma samples. **(C)** ROC and PR curve of a neural net (MLP) and AdaBoost classifier trained on the top 2000 highly variable genes in BCC and melanoma CD8$^+$ T cells respectively; both models were subsequently tested on each dataset separately. **(D, E)** Proportion of CD8$^+$ T cells classified as responsive to PD-1 blockade per patient by the **(D)** BCC neural net and **(E)** melanoma AdaBoost classifier.

2021), we find that PDAC suppresses both MHC-I and MHC-II expression in $CD8^+$ T cells even more severely by a factor of 2–5. This further reinforces prevailing beliefs that BCC and PDAC utilize divergent immune mechanisms in combating tumor progression. However, through our novel identification of malignant ductal cells in the PDAC TME, we find that the two cancers exhibit similar expression of MHC genes, although MHC-I and MHC-II expressions are slightly elevated in PDAC.

Strikingly, we were able to construct a classifier to predict response to PD-1 blockade in BCC $CD8^+$ T cells with near-perfect accuracy (97%). Even when considering data from only a handful of highly variable genes, responders and nonresponders were clearly distinguished. These results may suffer from overfitting due to the lack of suitable testing data: it is unknown whether the accuracy is artificially high due to the relative homogeneity of the TMEs of the 11 patients studied, or that the classifier will remain as successful in predicting the outcomes of other BCC cases. It is very likely that these models neglected to encompass the full spectrum of BCC subtypes, particularly nodular types with a high rate of heterogeneous morphological features both intra- and inter-tumorally (Pirie et al., 2018). However, our results strongly support that rapid and affordable testing of BCC patients, focusing on a small number (10–20) of genes, can accurately predict their response to checkpoint immunotherapies. Unfortunately, such clean results were not attained when training classifiers on the melanoma dataset or when testing either on PDAC; in these cases, our classifiers did not perform significantly better than random chance, mirroring previous efforts in the field (Banchereau et al., 2021).

The dearth of clinical data on the positive response of PDAC patients to PD-1 blockade leaves open multiple avenues of exploration. While the present study offers convincing evidence that PDAC, BCC, and melanoma lie in unique positions on the spectrum of immunogenicity, little is known of the exact changes in the immune landscape triggered by PD-1 blockade treatment, even in responders. A recent study on the same BCC dataset focusing on cell-cell communications offers several possible avenues of investigation, including the role of multiple tumor necrosis factor (*TNF*) pathways and a unique subtype of $CD8^+$ T cells characterized by high expression of suppressive, cytotoxic, and heat shock protein genes (Jiang et al., 2021). Furthermore, there is a crucial need for further research into actionable mechanisms to overcome resistance to immune checkpoint inhibitors – current studies point to the potentiality of combination therapies in delivering individualized, multi-faceted remodeling of the TME (Ott et al., 2017) (Drake, 2012). Other hypothesized factors for immune resistance include tumor exploitation of the PD-1/PD-L axis, immuno-editing in tumor cells, the immunosuppressive effects of long non-coding RNAs (lncRNAs), and insufficient re-invigoration of exhausted $CD8^+$ T cells (Drake et al., 2006) (Sun et al., 2020).

We chose to focus on the role of MHC-I and MHC-II expression in this investigation due to its well-established role in stimulating immune responses, as well as the obvious choice of PD-1/PD-L. However, various other genes and pathways associated with resistance to PD-1 blockade, such as *LAG-3* and the *IDO* pathway, were not studied in-depth (LaFleur et al., 2018) (Chocarro de Erauso et al., 2020). Additionally, recent studies have suggested that M2 macrophages and memory B cells play vital roles in directly affecting cancer cells (Dollinger et al., 2020) (Drake et al., 2006). With only six responders and five nonresponders in our BCC dataset, it is also likely that many cancer subtypes and diverse response mechanisms were not detected.

Despite the purely computational nature of the present study, our novel attempt to carry out quantitative comparisons of completely different cancers using scRNA-seq will facilitate a greater understanding of the immune landscape through the identification of both differences and similarities across different TMEs. Fundamentally, studying the activities of the same celltype in different TMEs serves an equivalent purpose to investigating the role of the same genes in different cancers. Although BCC and PDAC reside on opposite extremes of the spectrum of immunogenicity, the parallels that can be drawn between them will point the way towards establishing new immuno-oncology paradigms for more personalized and sophisticated immunotherapies.

4 MATERIALS AND METHODS

4.1 Clustering

All clustering analyses were performed using Seurat (version 3.6) (Stuart et al., 2019). The UMI matrices for the BCC dataset (Yost et al., 2019) and PDAC dataset (Steele et al., 2020) were downloaded from GEO accession GSE123813 and GSE155698 respectively; the count matrix for the melanoma dataset (Sade-Feldman et al., 2018) was personally contributed by the authors of (Dollinger et al., 2020). No clinical trials were performed in the data acquisition or any other part of the preparation of this paper. The following procedures were applied to both the BCC and PDAC dataset; preparation and analysis of the melanoma dataset is separately dealt with in **Section 4.4**.

To exclude low-quality cells and empty droplets, we excluded all cells with less than 200 features detected; furthermore, we excluded all features that were not present in at least three cells. In preparation for clustering, we then followed the preprocessing steps detailed in (Stuart et al., 2019). Briefly, we first normalized the feature expression using the Seurat LogNormalization method with defaults, then applied linear transformation to shift the mean expression of each gene to 0 and the variance to 1. We then identified the top 2000 highly variable genes through calculating the "standardized variance" of each feature, which captures single-cell dispersion in the context of mean expression. Using these features, linear dimensional reduction was conducted on the normalized through PCA. The first 50 PCs for the BCC dataset and first 20 PCs for the PDAC dataset were used to construct a KNN graph; the number of PCs used was determined using the Seurat `ElbowPlot` function by identifying the cutoff at which the percentage of variance explained by each additional PC dropped significantly. The Louvian algorithm was then applied on the KNN graph to group cells together. The "granularity" of the clustering was

determined by a resolution parameter which was set to 0.4 for BCC and 0.5 for PDAC (Seurat default = 0.3, higher resolutions correspond with a greater number of clusters).

To identify clusters, differential expression was performed to identify the top upregulated features in each cluster in comparison to all other clusters. Features were considered to be upregulated if at least 25% of cells in the cluster expressed the gene, the mean expression was greater by at least a factor of $2^{0.25}$, and the *p*-value was less than 0.05. The top 25 DEGs were then recorded and entered into Enrichr for gene enrichment analysis (Xie et al., 2021). Using a combination of different gene datasets (e.g. Human Gene Atlas and Mouse Gene Atlas) and common celltype marker genes, clusters were holistically identified. To ensure that equally-named clusters between BCC and PDAC (e.g. $CD4^+$ T cells) were identified similarly and were suitable for downstream comparison, cluster identification was checked using marker genes identified in both (Yost et al., 2019) and (Steele et al., 2020).

To validate the results of our clustering of the BCC dataset, we constructed a heatmap to compare our cell labels with those provided in the metadata of GEO accession GSE123813 (**Supplementary Figure S2A**). Specifically, we calculated the percentage of cells in each metadata cluster that was classified into each self-identified cluster. As no celltype identification was supplied in the PDAC dataset, we were unable to do the same for this dataset.

4.2 Statistical Analyses of Datasets

All statistical analyses were performed in RStudio version 1.4. We merged the BCC and PDAC datasets together without batch correction to compare expression of key marker genes in each cluster (**Figure 1F**) and relative population sizes of each cluster (**Figure 1G**); integration was not necessary as no new clustering was conducted. To calculate the breakdown of each cluster in the merged BCC + PDAC dataset between BCC responders, BCC nonresponders, PDAC cancerous samples, and PDAC adjacent samples (**Figure 1G**), we first normalized the total size of each of the four batches so that each batch would appear to have the same total number of cells. We then determined the normalized proportion of cells in each cluster that belonged to each batch by dividing the normalized population size of each batch in each cluster by the total normalized population of the cluster.

To compare the proportion of all cells that were identified as T cells (**Figure 2B**), we first determined the number of cells in each patient sample. Then, we calculated the fraction of these cells that were labeled as either $CD4^+$, $CD8^+$ effector, $CD8^+$ memory, $CD8^+$ exhausted, regulatory (Tregs), proliferating, NK, or miscellaneous T cells for each sample. To determine whether differences in this percentage between different groups were statistically significant, we performed Wilcoxon tests using the stat_compare_means function in the ggpubr package, version 0.4.0. A *p*-value lower than 0.05 was considered to be statistically significant. The same procedure was repeated for **Figure 2C, D**.

Comparisons of the distribution of expression of any particular gene between different clusters and/or categories (**Figures 2F–I**) were conducted by first taking the data from the normalized UMI matrix, then exponentiating all of the values so that any comparisons occur in non-log space. The distribution of these values were plotted using the ggviolin function in the ggpubr package; statistical significance was determined through a Wilcoxon test. Identical procedures were used in **Figures 3C–F**. The same method was also used to determine the *p*-values in **Figure 2J**; however, the log2-fold difference was calculated by dividing the mean expression of the particular gene of all BCC cells to the mean expression of all PDAC cells in the particular cluster. Mean expression was calculated using the AverageExpression function in Seurat and was therefore performed on raw data counts, as opposed to scaled/normalized data.

To perform full differential gene expression between two clusters (**Figure 1E**, **Figure 2E**, and **Figure 3B**), the Seurat objects of the clusters of interest were first merged together. Then, the EnhancedVolcano function from the Bioconductor package was used to generate the volcano plots in **Figure 1E** and **Figure 3B**. Features that were considered to be differentially expressed were those with a *p*-value <0.05 and a log2-fold absolute change greater than 0.5. For **Figure 2E**, the log2-fold change for every gene was ordered and normalized ($\mu = 0$, $\sigma = 1$); then, the 2000 genes with the lowest absolute value normalized log2-fold change were identified as the genes with most similar expression.

Heatmaps of gene expression were generated using the DoHeatmap function from Seurat. The genes displayed are the top *n* DEGs per cluster ($n = 3$ and 25 for **Figure 1F** and **Figure 3A** respectively).

4.3 Inference of Intercellular Communication Network Strengths

Cell-cell communication was determined using CellChat version 1.1 (Jin et al., 2021). Briefly, the cell-cell communication network was inferred by calculating the interaction probabilities, which is directly dependent on average gene expression, for each ligand-receptor pair in the CellChat database. The sum of communication probabilities of outgoing signaling from and incoming signaling to a particular cluster determines its outgoing and incoming interaction strength respectively, as plotted in **Figure 2K**.

4.4 Supervised Learning: Prediction of Response to PD-1 Blockade in BCC and Melanoma

Classifiers were constructed on $CD8^+$ T cells in BCC and melanoma (**Figure 4**). The BCC dataset consisted of all pre-treatment cells identified as $CD8^+$ effector, $CD8^+$ memory, or $CD8^+$ exhausted. The melanoma dataset consisted of all $CD8^+$ T cells as identified in (Dollinger et al., 2020).

All machine learning was conducted in Python using the scikit-learn package (Pedregosa et al., 2011). Identification of the top 2000 highly variable genes (**Figures 4A, B**) recapitulated the process described in **Section 4.1**. To determine the best model

in differentiating cells from responders and nonresponders, nine separate classifiers were trained separately on the BCC and melanoma datasets:

- Nearest Neighbors (sklearn.neighbors.KNeighborsClassifier): three neighbors
- Linear SVM (sklearn.svm.SVC): linear kernal, $C = 0.025$
- RBF SVM (sklearn.svm.SVC): $\gamma = 2$, $C = 1$
- Gaussian Process (sklearn.Gaussian_process.GaussianProcess Classifier)
- Decision Tree (sklearn.tree.DecisionTreeClassifier): max_depth = 5
- Random Forest (sklearn.ensemble.RandomForestClassifier): max_depth = 5, max_estimators = 10, max_features = 1
- Neural Net (sklearn.neural_network.MLPClassifier): $\alpha = 1$, max_iterations = 1000
- AdaBoost (sklearn.ensemble.AdaBoostClassifier)
- Naive Bayes (sklearn.naive_bayes.GaussianNB)
- Quadratic Classifier (sklearn.discriminant_analysis. QuadraticDiscriminantAnalysis)

All parameters used are the default ones unless listed explicitly. Models were trained on 80% of the dataset and tested on the remaining 20%. During parameter optimization, the mean score from five-fold cross validation was used to evaluate accuracy. The best classifier was chosen as the one with the highest accuracy, i.e. proportion of true positives and true negatives. For BCC $CD8^+$ T cells, this was the neural net classifier; for melanoma $CD8^+$ T cells, this was the AdaBoost classifier. After parameter optimization, the BCC classifier had an architecture of one hidden layer with 20 nodes, a rectified linear unit (relu) activation function ($f(x) = \max(0, x)$), and a stochastic-gradient based optimizer (adam); the learning rate is $\alpha = 1$ and all other hyperparameters are equal to function defaults. The melanoma classifier has an architecture of 500 estimators using the SAMME. R real boosting algorithm and a learning rate of $\alpha = 1$. To ensure consistency, all classifiers trained using a reduced number of highly variable features (**Supplementary Figure S4A, B**) used the same architectures.

To calculate the proportion of cells in each patient that are classified as responsive (**Figures 4D, E**), the BCC neural net classifier was tested on the BCC pretreatment and PDAC datasets, and the melanoma AdaBoost classifier was tested on the melanoma and PDAC datasets. The number of $CD8^+$ T cells classified as responsive in each patient was then divided by the total number of $CD8^+$ T cells in each patient. To determine whether the results were statistically significant, a Wilcoxon test was performed using the stat_compare_means function in ggpubr; p-values less than 0.05 were considered as significant.

DATA AVAILABILITY STATEMENT

Only publicly available datasets were analyzed in this study. This data can be found here: GEO accession GSE123813 for BCC and GSE155698 for PDAC. The melanoma dataset was directly provided by the authors of Sade-Feldman et al. and is available upon request. All code generated in preparing this article is publicly available at the following GitHub repository: https://github.com/rliu7926/bcc-pdac-pd1-blockade.

AUTHOR CONTRIBUTIONS

RL: Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review and Editing, Visualization. ED: Conceptualization, Methodology, Writing - Original Draft, Writing - Review and Editing, Supervision, Project Administration. QN: Conceptualization, Resources, Writing - Review and Editing, Supervision, Project Administration, Funding Acquisition.

FUNDING

This study was partially supported by NSF grant DMS1763272, Simons Foundation grant 594598, and NIH grant U54-CA217378. ED is supported by NIH grant T32GM136624.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Scott Atwood and the Atwood lab for thoughtful discussion of the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.806457/full#supplementary-material

Supplementary Figure S1 | Visual abstract. **(A)** BCC and PDAC dataasets are clustered and labeled through Seurat,resulting in the novel identification of malignant ductal cells in PDAC (**Section 2.1**). **(B)** Several analyseswere performed on T cells and malignant cells from both datasets, focusing on population subcluster sizes,differential gene expression, and cell-cell signaling (**Section 2.2**/**Section 2.3**). **(C)** Machine learning classification models were successfully utilized to predict whether individual CD8+ T cells in BCC and PDAC would respond to PD-1 blockade; however, these models are not transferrable onto PDAC (**Section 2.4**).

Supplementary Figure S2 | Validation of clustering in BCC and PDAC datasets. **(A)** Heatmap depicting the proportion of cells in each cluster of the original paper belonging to each cluster defined in this paper. **(B, C)** Dot plot of the expression of **(B)** CD8+ T cell and **(C)** ductal cell marker genes in PDAC. Color denotes the average expression across all cells in a subcluster, scaled per gene; size denotes the percentage of cells with positive expression within the subcluster.

Supplementary Figure S3 | MHC expression in malignant cells of BCC and PDAC. **(A, B)** Breakdown of **(A)** PDAC ductal cell and **(B)** BCC malignant cell clusters by patient. **(C, D)** Paired comparison of MHC-I and MHC-II scores per patient in **(C)** malignant vs. non-malignant PDAC ductal cells and **(D)** pre-treatment vs. post-treatment BCC malignant cells. Horizontal boxplots represent the log2-fold difference in the MHC score per patient between the two batches; the T-test calculates the likelihood that on average, there is no difference in the score.

Supplementary Figure S4 | Supervised classification of CD8+ T cells in BCC and melanoma based on the top *n* highlyvariable genes. **(A, B)** ROC and PR curves for classifier (architecture described in Methods) trained on the top n highly variable genes in **(A)** BCC and **(B)** CD8+ T cells. **(C)** Comparison of the proportion of top *n* highly variable genes that are common between BCC and melanoma CD8+ T cells. **(D)** Fold change differential expression of MHC and HSP genes between BCC and melanoma CD8+ T cells; positive values indicate greater expression in BCC.

REFERENCES

Arumugam, T., Brandt, W., Ramachandran, V., Moore, T. T., Wang, H., May, F. E., et al. (2011). Trefoil Factor 1 Stimulates Both Pancreatic Cancer and Stellate Cells and Increases Metastasis. *Pancreas* 40, 815–822. doi:10.1097/MPA.0b013e31821f6927

Bai, R., Lv, Z., Xu, D., and Cui, J. (2020). Predictive Biomarkers for Cancer Immunotherapy with Immune Checkpoint Inhibitors. *Biomark Res.* 8, 34. doi:10.1186/s40364-020-00209-0

Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A. M., Gingras, M. C., et al. (2016). Genomic Analyses Identify Molecular Subtypes of Pancreatic Cancer. *Nature* 531, 47–52. doi:10.1038/nature16965

Banchereau, R., Leng, N., Zill, O., Sokol, E., Liu, G., Pavlick, D., et al. (2021). Molecular Determinants of Response to PD-L1 Blockade across Tumor Types. *Nat. Commun.* 12, 3969. doi:10.1038/s41467-021-24112-w

Bengtsson, A., Andersson, R., and Ansari, D. (2020). The Actual 5-year Survivors of Pancreatic Ductal Adenocarcinoma Based on Real-World Data. *Sci. Rep.* 10, 16425. doi:10.1038/s41598-020-73525-y

Bonilla, X., Parmentier, L., King, B., Bezrukov, F., Kaya, G., Zoete, V., et al. (2016). Genomic Analysis Identifies New Drivers and Progression Pathways in Skin Basal Cell Carcinoma. *Nat. Genet.* 48, 398–406. doi:10.1038/ng.3525

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., et al. (2017). Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational burden. *Genome Med.* 9, 34. doi:10.1186/s13073-017-0424-2

Chocarro de Erauso, L., Zuazo, M., Arasanz, H., Bocanegra, A., Hernandez, C., Fernandez, G., et al. (2020). Resistance to Pd-L1/pd-1 Blockade Immunotherapy. A Tumor-Intrinsic or Tumor-Extrinsic Phenomenon? *Front. Pharmacol.* 11, 441. doi:10.3389/fphar.2020.00441

Christenson, E. S., Jaffee, E., and Azad, N. S. (2020). Current and Emerging Therapies for Patients with Advanced Pancreatic Ductal Adenocarcinoma: a Bright Future. *Lancet Oncol.* 21, e135–e145. doi:10.1016/s1470-2045(19)30795-8

Ciocca, D. R., and Calderwood, S. K. (2005). Heat Shock Proteins in Cancer: Diagnostic, Prognostic, Predictive, and Treatment Implications. *Cell Stress Chaper* 10, 86. doi:10.1379/csc-99r.1

Dhatchinamoorthy, K., Colbert, J. D., and Rock, K. L. (2021). Cancer Immune Evasion through Loss of Mhc Class I Antigen Presentation. *Front. Immunol.* 12, 469. doi:10.3389/fimmu.2021.636568

Dolezal, J. M., Dash, A. P., and Prochownik, E. V. (2018). Diagnostic and Prognostic Implications of Ribosomal Protein Transcript Expression Patterns in Human Cancers. *BMC Cancer* 18, 275. doi:10.1186/s12885-018-4178-z

Dollinger, E., Bergman, D., Zhou, P., Atwood, S. X., and Nie, Q. (2020). Divergent Resistance Mechanisms to Immunotherapy Explain Responses in Different Skin Cancers. *Cancers (Basel)* 12, 946. doi:10.3390/cancers12102946

Drake, C. G. (2012). Combination Immunotherapy Approaches. *Ann. Oncol.* 23, viii41–viii46. doi:10.1093/annonc/mds262

Drake, C. G., Jaffee, E., and Pardoll, D. M. (2006). "Mechanisms of Immune Evasion by Tumors," in *Cancer Immunotherapy. Vol. 90 of Advances in Immunology* (Cambridge, Massachusetts, USA: Academic Press), 51–81. doi:10.1016/S0065-2776(06)90002-9

Fan, J.-q., Wang, M.-F., Chen, H.-L., Shang, D., Das, J. K., and Song, J. (2020). Current Advances and Outlooks in Immunotherapy for Pancreatic Ductal Adenocarcinoma. *Mol. Cancer* 19, 32. doi:10.1186/s12943-020-01151-3

Foucher, E. D., Ghigo, C., Chouaib, S., Galon, J., Iovanna, J., and Olive, D. (2018). Pancreatic Ductal Adenocarcinoma: A strong Imbalance of Good and Bad Immunological Cops in the Tumor Microenvironment. *Front. Immunol.* 9, 1044. doi:10.3389/fimmu.2018.01044

Grund-Gröschke, S., Ortner, D., Szenes-Nagy, A. B., Zaborsky, N., Weiss, R., Neureiter, D., et al. (2020). Epidermal Activation of Hedgehog Signaling Establishes an Immunosuppressive Microenvironment in Basal Cell Carcinoma by Modulating Skin Immunity. *Mol. Oncol.* 14, 1930–1946. doi:10.1002/1878-0261.12758

Grzywa, T. M., Paskal, W., and Włodarski, P. K. (2017). Intratumor and Intertumor Heterogeneity in Melanoma. *Translational Oncol.* 10, 956–975. doi:10.1016/j.tranon.2017.09.007

Jiang, Y.-Q., Wang, Z.-X., Zhong, M., Shen, L.-J., Han, X., Zou, X., et al. (2021). Investigating Mechanisms of Response or Resistance to Immune Checkpoint Inhibitors by Analyzing Cell-Cell Communications in Tumors before and after Programmed Cell Death-1 (PD-1) Targeted Therapy: An Integrative Analysis Using Single-Cell RNA and Bulk-RNA Sequencing Data. *OncoImmunology* 10, 1908010. doi:10.1080/2162402x.2021.1908010

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., et al. (2021). Inference and Analysis of Cell-Cell Communication Using Cellchat. *Nat. Commun.* 12, 1088. doi:10.1038/s41467-021-21246-9

Johnson, D. B., Estrada, M. V., Salgado, R., Sanchez, V., Doxie, D. B., Opalenik, S. R., et al. (2016). Melanoma-specific MHC-II Expression Represents a Tumour-Autonomous Phenotype and Predicts Response to Anti-PD-1/pd-l1 Therapy. *Nat. Commun.* 7, 10582. doi:10.1038/ncomms10582

Kunovsky, L., Tesarikova, P., Kala, Z., Kroupa, R., Kysela, P., Dolina, J., et al. (2018). The Use of Biomarkers in Early Diagnostics of Pancreatic Cancer. *Can. J. Gastroenterol. Hepatol.* 2018, 1–10. doi:10.1155/2018/5389820

LaFleur, M. W., Muroyama, Y., Drake, C. G., and Sharpe, A. H. (2018). Inhibitors of the PD-1 Pathway in Tumor Therapy. *J. Immunol.* 200, 375–383. doi:10.4049/jimmunol.1701044

Leclerc, E., and Vetter, S. W. (2015). The Role of S100 Proteins and Their Receptor Rage in Pancreatic Cancer. *Biochim. Biophys. Acta* 1852, 2706–2711. doi:10.1016/j.bbadis.2015.09.022

Liu, Y., Liu, H., and Bian, Q. (2020). Identification of Potential Biomarkers Associated with Basal Cell Carcinoma. *Biomed. Res. Int.* 2020, 1–10. doi:10.1155/2020/2073690

Mehner, C., and Radisky, E. S. (2019). Bad Tumors Made Worse: Spink1. *Front. Cel Dev. Biol.* 7, 10. doi:10.3389/fcell.2019.00010

Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G. H., Hoadley, K. A., et al. (2015). Virtual Microdissection Identifies Distinct Tumor- and Stroma-specific Subtypes of Pancreatic Ductal Adenocarcinoma. *Nat. Genet.* 47, 1168–1178. doi:10.1038/ng.3398

Moujaess, E., Merhy, R., Kattan, J., Sarkis, A.-S., and Tomb, R. (2021). Immune Checkpoint Inhibitors for Advanced or Metastatic Basal Cell Carcinoma: How Much Evidence Do We Need? *Immunotherapy* 13, 1293–1304. doi:10.2217/imt-2021-0089

Orecchioni, M., Ghosheh, Y., Pramod, A. B., and Ley, K. (2019). Macrophage Polarization: Different Gene Signatures in M1(LPS+) vs. Classically and M2(LPS-) vs. Alternatively Activated Macrophages. *Front. Immunol.* 10, 1084. doi:10.3389/fimmu.2019.01084

Ott, P. A., Hodi, F. S., Kaufman, H. L., Wigginton, J. M., and Wolchok, J. D. (2017). Combination Immunotherapy: a Road Map. *J. Immunother. Cancer* 5, 16. doi:10.1186/s40425-017-0218-5

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in python. *J. Machine Learn. Res.* 12, 2825–2830.

Pellegrini, C., Maturo, M., Di Nardo, L., Ciciarelli, V., Gutiérrez García-Rodrigo, C., and Fargnoli, M. (2017). Understanding the Molecular Genetics of Basal Cell Carcinoma. *Int. J. Mol. Sci.* 18, 2485. doi:10.3390/ijms18112485

Pirie, K., Beral, V., Heath, A. K., Green, J., Reeves, G. K., Peto, R., et al. (2018). Heterogeneous Relationships of Squamous and Basal Cell Carcinomas of the Skin with Smoking: the UK Million Women Study and Meta-Analysis of Prospective Studies. *Br. J. Cancer* 119, 114–120. doi:10.1038/s41416-018-0105-y

Possick, J. D. (2017). Pulmonary Toxicities from Checkpoint Immunotherapy for Malignancy. *Clin. Chest Med.* 38, 223–232. doi:10.1016/j.ccm.2016.12.012

Pu, N., Lou, W., and Yu, J. (2019). Pd-1 Immunotherapy in Pancreatic Cancer: Current Status. *J. Pancreatology* 2, 10. doi:10.1097/jp9.0000000000000010

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting Cancer Incidence and Deaths to 2030: The Unexpected burden of Thyroid, Liver, and Pancreas Cancers in the united states. *Cancer Res.* 74, 2913–2921. doi:10.1158/0008-5472.CAN-14-0155

Renouf, D. J., Knox, J. J., Kavan, P., Jonker, D., Welch, S., Couture, F., et al. (2020). Lba65 the canadian Cancer Trials Group pa.7 Trial: Results of a Randomized Phase Ii Study of Gemcitabine (Gem) and Nab-Paclitaxel (Nab-p) vs Gem, Nab-P, Durvalumab (D) and Tremelimumab (T) as First Line Therapy in Metastatic Pancreatic Ductal Adenocarcinoma (Mpdac). *Ann. Oncol.* 31, S1195. doi:10.1016/j.annonc.2020.08.2300

Reyes-Castellanos, G., Masoud, R., and Carrier, A. (2020). Mitochondrial Metabolism in Pdac: From Better Knowledge to New Targeting Strategies. *Biomedicines* 8, 270. doi:10.3390/biomedicines8080270

Ribas, A., and Wolchok, J. D. (2018). Cancer Immunotherapy Using Checkpoint Blockade. *Science* 359, 1350–1355. doi:10.1126/science.aar4060

Rodig, S. J., Gusenleitner, D., Jackson, D. G., Gjini, E., Giobbie-Hurder, A., Jin, C., et al. (2018). Mhc Proteins Confer Differential Sensitivity to Ctla-4 and Pd-1 Blockade in Untreated Metastatic Melanoma. *Sci. Transl. Med.* 10, eaar3342. doi:10.1126/scitranslmed.aar3342

Royal, R. E., Levy, C., Turner, K., Mathur, A., Hughes, M., Kammula, U. S., et al. (2010). Phase 2 Trial of Single Agent Ipilimumab (Anti-ctla-4) for Locally Advanced or Metastatic Pancreatic Adenocarcinoma. *J. Immunother.* 33, 828–833. doi:10.1097/CJI.0b013e3181eec14c

Sade-Feldman, M., Yizhak, K., Bjorgaard, S. L., Ray, J. P., de Boer, C. G., Jenkins, R. W., et al. (2018). Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* 175, 998–1013. e20. doi:10.1016/j.cell.2018.10.038

Šmahel, M. (2017). PD-1/PD-L1 Blockade Therapy for Tumors with Downregulated MHC Class I Expression. *Int. J. Mol. Sci.* 18, 1331. doi:10.3390/ijms18061331

Steele, N. G., Carpenter, E. S., Kemp, S. B., Sirihorachai, V. R., The, S., Delrosario, L., et al. (2020). Multimodal Mapping of the Tumor and Peripheral Blood Immune Landscape in Human Pancreatic Cancer. *Nat. Cancer* 1, 1097–1112. doi:10.1038/s43018-020-00121-4

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, et al. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888–1902. doi:10.1016/j.cell.2019.05.031

Sun, J.-Y., Zhang, D., Wu, S., Xu, M., Zhou, X., Lu, X.-J., et al. (2020). Resistance to PD-1/pd-L1 Blockade Cancer Immunotherapy: Mechanisms, Predictive Factors, and Future Perspectives. *Biomark Res.* 8, 35. doi:10.1186/s40364-020-00212-5

Tang, Y., Zhang, Z., Tang, Y., Chen, X., and Zhou, J. (2018). Identification of Potential Target Genes in Pancreatic Ductal Adenocarcinoma by Bioinformatics Analysis. *Oncol. Lett.* 16, 2453. doi:10.3892/ol.2018.8912

Tumeh, P. C., Harview, C. L., Yearley, J. H., Shintaku, I. P., Taylor, E. J. M., Robert, L., et al. (2014). Pd-1 Blockade Induces Responses by Inhibiting Adaptive Immune Resistance. *Nature* 515, 568–571. doi:10.1038/nature13954

Verma, V., Shrimali, R. K., Ahmad, S., Dai, W., Wang, H., Lu, S., et al. (2019). PD-1 Blockade in Subprimed CD8 Cells Induces Dysfunctional PD-1CD38hi Cells and Anti-PD-1 Resistance. *Nat. Immunol.* 20, 1231–1243. doi:10.1038/s41590-019-0441-y

Walter, A., Barysch, M. J., Behnke, S., Dziunycz, P., Schmid, B., Ritter, E., et al. (2010). Cancer-testis Antigens and Immunosurveillance in Human Cutaneous Squamous Cell and Basal Cell Carcinomas. *Clin. Cancer Res.* 16, 3562–3570. doi:10.1158/1078-0432.CCR-09-3136

Wang, Y., Tong, Z., Zhang, W., Zhang, W., Buzdin, A., Mu, X., et al. (2021). Fda-approved and Emerging Next Generation Predictive Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients. *Front. Oncol.* 11, 2115. doi:10.3389/fonc.2021.683419

Wolfgang, C. L., Herman, J. M., Laheru, D. A., Klein, A. P., Erdek, M. A., Fishman, E. K., et al. (2013). Recent Progress in Pancreatic Cancer. *CA A Cancer J. Clinicians* 63, 318–348. doi:10.3322/caac.21190

Wu, J., Liu, T., Rios, Z., Mei, Q., Lin, X., and Cao, S. (2017). Heat Shock Proteins and Cancer. *Trends Pharmacol. Sci.* 38, 226–256. doi:10.1016/j.tips.2016.11.009

Xia, A., Zhang, Y., Xu, J., Yin, T., and Lu, X.-J. (2019). T Cell Dysfunction in Cancer Immunity and Immunotherapy. *Front. Immunol.* 10, 1719. doi:10.3389/fimmu.2019.01719

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., et al. (2021). Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc.* 1, e90. doi:10.1002/cpz1.90

Yarchoan, M., Hopkins, A., and Jaffee, E. M. (2017). Tumor Mutational burden and Response Rate to Pd-1 Inhibition. *New Engl. J. Med.* 377, 2500–2501. doi:10.1056/NEJMc1713444

Yost, K. E., Satpathy, A. T., Wells, D. K., Qi, Y., Wang, C., Kageyama, R., et al. (2019). Clonal Replacement of Tumor-specific T Cells Following Pd-1 Blockade. *Nat. Med.* 25, 1251–1259. doi:10.1038/s41591-019-0522-3

Zhu, Y., Knolhoff, B. L., Meyer, M. A., Nywening, T. M., West, B. L., Luo, J., et al. (2014). Csf1/csf1r Blockade Reprograms Tumor-Infiltrating Macrophages and Improves Response to T-Cell Checkpoint Immunotherapy in Pancreatic Cancer Models. *Cancer Res.* 74, 5057–5069. doi:10.1158/0008-5472.CAN-13-3723

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 04 February 2022
doi: 10.3389/fgene.2022.825896

DecOT: Bulk Deconvolution With Optimal Transport Loss Using a Single-Cell Reference

Gan Liu[1], *Xiuqin Liu*[1]* *and Liang Ma*[2]*

[1]Department of Information and Computing Science, University of Science and Technology Beijing, Beijing, China, [2]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

Tissues are constituted of heterogeneous cell types. Although single-cell RNA sequencing has paved the way to a deeper understanding of organismal cellular composition, the high cost and technical noise have prevented its wide application. As an alternative, computational deconvolution of bulk tissues can be a cost-effective solution. In this study, we propose DecOT, a deconvolution method that uses the Wasserstein distance as a loss and applies scRNA-seq data as references to characterize the cell type composition from bulk tissue RNA-seq data. The Wasserstein loss in DecOT is able to utilize additional information from gene space. DecOT also applies an ensemble framework to integrate deconvolution results from multiple individuals' references to mitigate the individual/batch effect. By benchmarking DecOT with four recently proposed square loss-based methods on pseudo-bulk data from four different single-cell data sets and real pancreatic islet bulk samples, we show that DecOT outperforms other methods and the ensemble framework is robust to the choice of references.

Keywords: bulk RNA sequencing, single-cell RNA sequencing, cell-type deconvolution, wasserstein distance, optimal transport

OPEN ACCESS

Edited by:
Jiajun Zhang,
Sun Yat-sen University, China

Reviewed by:
Suoqin Jin,
Wuhan University, China
Xiaoqiang Sun,
Sun Yat-sen University, China

***Correspondence:**
Xiuqin Liu
mathlxq@163.com
Liang Ma
maliang@ioz.ac.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *30 November 2021*
Accepted: *04 January 2022*
Published: *04 February 2022*

Citation:
Liu G, Liu X and Ma L (2022) DecOT: Bulk Deconvolution With Optimal Transport Loss Using a Single-Cell Reference. Front. Genet. 13:825896. doi: 10.3389/fgene.2022.825896

INTRODUCTION

Quantification of gene expression changes in different tissues or under different conditions gives information on how genes are regulated in organisms. The analysis of gene expression by using RNA sequencing (RNA-seq) has contributed substantially, since its development more than a decade ago, to our understanding of biological processes such as organism development, human disease progression, and patients' response to treatments. The classic RNA-seq applied to bulk tissue samples has accumulated a rich reservoir of data sets, for example, GTEx, TCGA, and so forth (Tomczak et al., 2015, Carithers et al., 2015). However, since tissues are heterogeneous, which comprise a variety of cell types, the bulk sequencing data only measure the average state of the mixed cell populations. In fact, the information of cellular composition is crucial. For example, when developing diagnostic techniques, such information would enable researchers to track the contribution of each cellular component during disease progressions (Schelker et al., 2017).

With the rapid development of single-cell technologies, one way to obtain a cell-specific transcriptome is to apply single-cell RNA-seq (Saliba et al., 2014). However, these experiments remain costly and noisy compared to the mature bulk RNA-seq and have therefore been performed only on a limited scale (Denisenko et al., 2020): (Kuksin et al., 2021). Alternatively, one may apply computational deconvolution algorithms with bulk data, which provide cost-effective ways to derive

cellular composition information and have the potential to bring considerable improvements in the speed and scale of relevant applications.

In recent years, a number of computational deconvolution methods have been developed with the goal of estimating cell-type composition within the bulk sample and/or cell-type-specific states (Avila Cobos et al., 2018); (Jin and Liu, 2021). According to whether references, such as expression profiles of pure cell types or marker gene lists, are provided, these deconvolution methods can be divided into supervised and unsupervised categories. As completely unsupervised approaches based on non-negative matrix factorization (NMF) suffer from low deconvolution accuracy and interpretation of their results largely depends on the ability to recover meaningful gene features or expression profiles for different cell types, the most commonly used methods are under the supervised category and are often optimized by least squares algorithms (Avila Cobos et al., 2018). The rapid accumulation of publicly available scRNA-seq data on a number of different samples (Baron et al., 2016), (Guo, 2020), led to the recent popularization of developing deconvolution methods with scRNA-seq references. For instance, Bisque learned the gene-specific conversion of bulk data from the scRNA-seq reference, eliminating the technical deviation of the sequencing technology between reference and bulk data (Jew et al., 2020). MuSiC proposes a weighted non-negative least squares regression framework that simultaneously weighs each gene through cross-subject and cross-cell variation (Wang et al., 2019). SCDC extends the MuSiC method and proposes an ensemble framework which applies multiple scRNA-seq data sets as reference deconvolution. They claim that SCDC can implicitly solve the batch effect between reference data sets in different experiments (Dong et al., 2019).

Besides square loss, divergence functions for characterizing differences between two distributions, for example, Kullback-Leibler divergence, are also commonly applied as loss functions in solving deconvolution problems (Lee and Seung, 1999). These losses, as well as square losses, decompose vectors or distributions in an elementwise manner, which neglects relationships between features (in our case, correlations between genes) (Zhang, 2021), (Afshar et al., 2020).

Recently, the Wasserstein distance, which originated from the optimal transport (OT) problem (Monge, 1781); (Kantorovich, 1942), has shown its potential as a better loss function for measuring the distance between distributions (Langfelder and Horvath, 2008); (Arjovsky et al., 2017). Wasserstein distance utilizes a metric between features (e.g., genes) called ground cost to take advantage of additional knowledge from the feature space (Rolet et al., 2016). Especially, when comparing two non-overlapping distributions (distributions with non-overlapping support), Wasserstein distance can still provide a smooth and meaningful measure, which is a desirable property that square loss and other divergence losses cannot offer (Weng, 2019), (Schmitz et al., 2018a). Since the first application of Wasserstein loss in solving NMF problems in Sandler and Lindenbaum, 2011, it has been successfully applied to blind source decomposition (Rolet et al., 2018), dictionary learning (Rolet et al., 2016), (Schmitz et al., 2018b), and multilabel supervised learning problems.

Cell types are characterized in gene space. The expression of genes is not mutually independent. The co-expression of genes naturally induces a similarity or distance metric among genes (Langfelder and Horvath, 2008). To the best of our knowledge, such a relationship has not yet been leveraged to solve cell-type devolution problems.

Here, we present DecOT, a bulk gene expression deconvolution method that uses the optimal transport distance as a loss and applies an ensemble framework to integrate reference information from scRNA-seq data of multiple individuals. We apply different ground cost metrics for characterizing gene relations in DecOT. We optimize DecOT under an entropic regularized form. We test the performance of DecOT on pseudo-bulk mixtures generated from different data sets and evaluate its robustness when different reference data are supplied. Finally, we applied DecOT on a real pancreatic islet bulk data set. DecOT is available on GitHub (https://github.com/lg-ustb/DecOT).

MATERIALS AND METHODS

In this section, we will first give a brief review of the original Wasserstein distance and the optimization algorithm with entropic regularization. Then, we will introduce our proposed DecOT framework for deconvolution. Finally, we will describe the data sets and procedures used for benchmarking DecOT.

Wasserstein Distance and Entropic Regularization

Wasserstein distance, originated from the optimal transport problem (Monge, 1781); (Kantorovich, 1942), aims at minimizing transportation costs between two probability distributions. Given two histograms, $p \in \Sigma_n$ and $q \in \Sigma_s$, the Wasserstein distance between p and q with respect to ground cost M is

$$W(p,q)_M \overset{\text{def}}{=} \min_{T \in U(p,q)} < M, T > \tag{1}$$

where $\Sigma_n \overset{\text{def}}{=} \{q \epsilon \mathbb{R}_+^n | < q, 1 > = 1\}$ is the set of histograms or an n-dimensional simplex; $< X, Y > \overset{\text{def}}{=} tr(X^T Y) = \sum_{i=1}^m X_i, Y_i$ is the Frobenius dot product between matrices X and Y; $U(p,q) = \left\{ T \in \mathbb{R}_+^{n \times s} | \begin{matrix} T1 = p \\ T^T 1 = p \end{matrix} \right\}$ is called the transportation polytope of p and q; M is the transportation cost of mapping p to q, which is also called the ground cost. W is a distance whenever M_{ij} is a metric in these two histograms' element space (Villani, 2009).

The computation of Wasserstein distance is extremely costly when the histograms' dimension exceeds a few hundreds. Cuturi et al. (Cuturi, 2013) introduced an entropic regularizer to smooth the optimal transport problem, which can be computed at several orders of a magnitude faster in speed than traditional algorithms

$$W_\gamma(p,q)_M \overset{\text{def}}{=} \min_{T\in U(p,q)} < M, T > - \gamma h(T) \quad (2)$$

where $\gamma > 0$ is a hyperparameter. $h(T) \overset{\text{def}}{=} - < T, \log T > = -\sum_{i,j} T_{ij}\log(T_{ij})$ is the entropic function.

The problem **(Eq. 2)** is strongly convex, and the solution of transport plan T^* can be optimized by solving a matrix balancing problem, which is typically solved using the fixed point Sinkhorn algorithm (Sinkhorn, 1967). The hyperparameter γ plays an important role in the final performance of Sinkhorn, with higher values of γ corresponding to a faster execution of the algorithm but a more diffused coupling. In this study, unless otherwise noted, we use $\gamma = 0.001$ by default.

Cell-Type Deconvolution with Wasserstein Loss

In this section, we will introduce the bulk tissue deconvolution framework by applying the Wasserstein distance as a loss function, which is the core part of DecOT.

We assume that each cell type has a unique expression profile which can be characterized by a distribution/histogram in gene space; for instance, we denote the expression profile over n genes of cell type i as $C_i \in \Sigma_n$. Thus, the cell type-specific profiles of k types can be represented as a $k \times n$ matrix,$C \in \Sigma_n^k$. For a set of normalized bulk tissue samples $Y = \{Y_1, \ldots, Y_m : Y_j \in \Sigma_n, \forall j\}$, the deconvolution problem is to solve the cell-type proportion or mixture proportion $P \in \Sigma_k^m$ for the m bulk samples by giving cell-type-specific profiles C, which can be represented by

$$Y \approx C \cdot P$$

To avoid individual/batch effects, here, we use reference data from a single individual. The annotated scRNA-seq reference data are then used by averaging the cell expressions within each cell type to generate C. The Wasserstein distance not only measures the difference between two distributions but also accounts for the underlying geometry of the feature (gene) space through the choice of an appropriate ground cost. Since the expression of genes is not mutually independent, the co-expression pattern between pairs of genes naturally induces a similarity or a distance metric among genes. Such a relationship forms the transportation cost among genes (ground cost M) and will be incorporated in the minimization of Wasserstein distance between the bulk sample gene expression distribution Y and the estimated mixture $C\hat{P}$. In order to ensure a trackable calculation for data containing thousands of genes, we apply the entropic regularized Wasserstein distance as a loss, which results in solving the following optimization problem

$$\min_{P\in\Sigma_k^m} \sum_{j=1}^m W_\gamma\left(Y_j, CP_j\right)_M \; s.t.\, CP \in \Sigma_n^m \quad (3)$$

In addition, since the cell-type proportions are non-negative, we further added a regularization term, as performed by Rolet et al. (Rolet et al., 2016) in solving the dictionary learning problem with a fixed dictionary, to enforce non-negativity constraints on the variables

$$\min_{P\in\Sigma_k^m} \sum_{j=1}^m W_\gamma\left(Y_j, CP_j\right)_M - \rho E\left(P_j\right) \; s.t.\, CP \in \Sigma_n^m \quad (4)$$

where E is defined for matrices whose columns are in the simplex as $E(A) = < A, \log A >$ and $\rho > 0$ is a hyperparameter. In this study, unless otherwise noted, we use $\rho = 0.001$ by default.

Ensemble Deconvolution Results Across Individuals

With the accumulation of publicly available single-cell data, references from multiple individuals may be available. In order to resolve variabilities in gene expression between references from different individuals, we adopt an ensemble approach similar to SCDC (Dong et al., 2019). The difference is that we focus on individuals rather than reference data sets of different experimental platforms. Assuming that single-cell data sets from R reference individuals are available, we first deconvolve the bulk gene expression data with entropic regularized Wasserstein loss as described above for each individual reference. Let $\hat{C}_{(r)}$ and $\hat{P}_{(r)}$ denote the cell-type-specific average expression matrix and the cell-type proportion matrix computed from the r^{th} reference individual. Our goal is to find the optimal combination strategy to ensemble the available deconvolution results

$$(\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_R) = \underset{(w_1,w_2,\ldots,w_R)}{\arg\min}\; l(P, \sum_{r=1}^R w_r \hat{P}_{(r)}) \quad (5)$$

where l is the loss function.

As explained by Dong in SCDC (Dong et al., 2019), function (5) cannot be optimized directly since the actual cell-type proportions P are unknown, and the solutions to function (5) are approximately equivalent to minimize the loss of gene expression levels. Therefore, we change the optimization problem to

$$(\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_R) = \underset{(w_1,w_2,\ldots,w_R)}{\operatorname{argmin}}\; l(Y, \sum_{r=1}^R w_r \hat{Y}_{(r)})$$

where $\hat{Y}_{(r)} = \hat{C}_{(r)}\hat{P}_{(r)}$ is the r^{th} individual's predicted bulk gene expression levels.

We redefine the problem to non-negative least squares regression by choosing the l_2 norm as loss

$$\min \left\| Y - \sum_{r=1}^R w_r \hat{Y}_{(r)} \right\|_2 \; s.t. \sum_{r=1}^R w_r = 1,\; w_r > 0$$

Intuitively, w_r can be seen as the similarity of cell expression profiles between r^{th} reference individual and a bulk tissue-derived individual.

Ground Cost Selection

In Wasserstein distance, a key factor is the ground cost matrix M, which defines the transportation cost. We obtain M from the reference cells an expression histogram X whose columns correspond to cells and whose rows correspond to genes. M_{ij} represents the dissimilarity of expression between gene i and

TABLE 1 | Four real scRNA-seq data sets.

Data set	Tissue type	Data type	Protocol	Individual samples	Cells	Genes	Cell types
Baron (GSE84133) Baron et al. (2016)	Pancreatic islet	Single-cell RNA-seq	Illumina HiSeq 2,500 (InDrop)	4	7,876	8,415	10
E-MTAB-5061 Segerstolpe et al. (2016)	Pancreatic islet	Single-cell RNA-seq	Smart-seq2	10	1901	14,200	7
GSE81547 Enge et al. (2017)	Pancreatic islet	Single-cell RNA-seq	Smart-seq2	8	2073	11,861	5
Kidney.HCL Han et al. (2020) Guo, (2020)	Kidney	Single-cell RNA-seq	Microwell-seq	3	20,601	2,748	13

gene j in reference cells. Here, we focus on four metrics, including

(i) Euclidean distance: $||x-y||_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.
(ii) Cosine similarity: $\cos(x,y) = \frac{xy}{||x||_2 \times ||y||_2}$. We use $1 - \cos(x,y)$ as distance.
(iii) Pearson correlation: $\mathrm{cor}(x,y) = \cos(x-\bar{x}, y-\bar{y})$, where $\bar{x}$ and $\bar{y}$ are the mean of the values of x and y, respectively. We use $1 - \mathrm{cor}(x,y)$ as distance.
(iv) Topological overlap-based dissimilarity measure (dissTOM) (Ravasz et al., 2002; Li and Horvath, 2007; Yip and Horvath, 2007) underweighted gene co-expression network analysis framework (Zhang and Horvath, 2005)

$$d_{ij}^{\omega} = 1 - \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{\min\left\{\sum_u a_{iu}, \sum_u a_{ju}\right\} + 1 - a_{ij}}$$

where a_{ij} is the power adjacency function. dissTOM metric measures the distance between genes in a co-expression network, which is converted into a scale-free network. We use a python package POT (Flamary et al., 2021) to compute metrics (i)–(iii) and WGCNA (Langfelder and Horvath, 2008)..., a R package ... to compute dissTOM.

Benchmark Data Sets and Artificial Pseudo-bulk Mixtures

To evaluate DecOT and compare it to other deconvolution methods using l_2 norm loss, we generated artificial pseudo-bulk mixtures from four real RNA-seq data sets (see **Table 1**). We partly adopt the preprocessing and quality control pipeline in Cobos et al. (Avila Cobos et al., 2020) to the original data, which include filtering genes with all zero expression or zero variance, removing cells with the library size deviating from the mean size over three median absolute deviations (MADs), keeping genes with at least 5% of all cells having a UMI or read count greater than 1, and retaining cell types with at least 50 cells passing the quality control step (Avila Cobos et al., 2020).

After quality control, for each individual in each data set, we split their cells evenly into the reference set and testing set with similar distribution of cell types. Then, we generate 200 pseudo-bulk mixtures by randomly sampling 60% of the cells each time in testing data sets and aggregate the expression counts of each gene to generate the pseudo-bulk sample. The true cell-type proportions are recorded, which allows us to benchmark the performance of different deconvolution methods. The flow chart for constructing pseudo-bulk mixtures is shown in **Supplementary Figure S1**.

To evaluate the performance of deconvolution methods, we need to measure the deviation of the estimated proportion $\hat{P}$ to the true P. Here, we apply the Pearson correlation coefficient and root-mean-squared error (RMSE) to evaluate the performance of deconvolution methods:

(i) Pearson correlation: $cor(P, \hat{P})$;
(ii) Root-mean-squared error: RMSE $= \sqrt{\frac{1}{k \cdot m}\sum_i^k \sum_j^m (P_{i,j} - \hat{P}_{i,j})^2}$.

RESULTS

Method Overview

Since Wasserstein distance has been successfully applied to blind source decomposition (Rolet et al., 2018) and dictionary learning (Rolet et al., 2016), (Schmitz et al., 2018b) problems with excellent performance, we aimed to apply Wasserstein loss on the bulk deconvolution problem. We propose DecOT, which applies Wasserstein loss to estimate the relative abundance of cell types within a bulk sample by using a scRNA-seq reference ensemble of multi-individuals. An overview of DecOT is shown in **Figure 1**. DecOT first solves the entropic regularized Wasserstein loss for the cell-type deconvolution problem (*Cell Type Deconvolution with Wasserstein Loss* **formula 4**) based on a single individual reference constitute of scRNA-seq data with annotated cell types. Wasserstein distance aims to find the optimal transport plan under a given transportation cost. In our case, the transportation cost, also referred to as the "ground cost," represents the similarity or distance among genes. Therefore, the application of Wasserstein loss can take advantage of the relationship between genes to get an accurate estimate.

When references from multi-individuals are available, to minimize the possible bias induced by individual and/or platform variations across different individual references, we apply an ensemble framework similar to SCDC (Dong et al., 2019), which aims to solve batch effects between reference data sets. Instead of weighting deconvolution results across a data set, DecOT seeks to optimize weights on results based on each

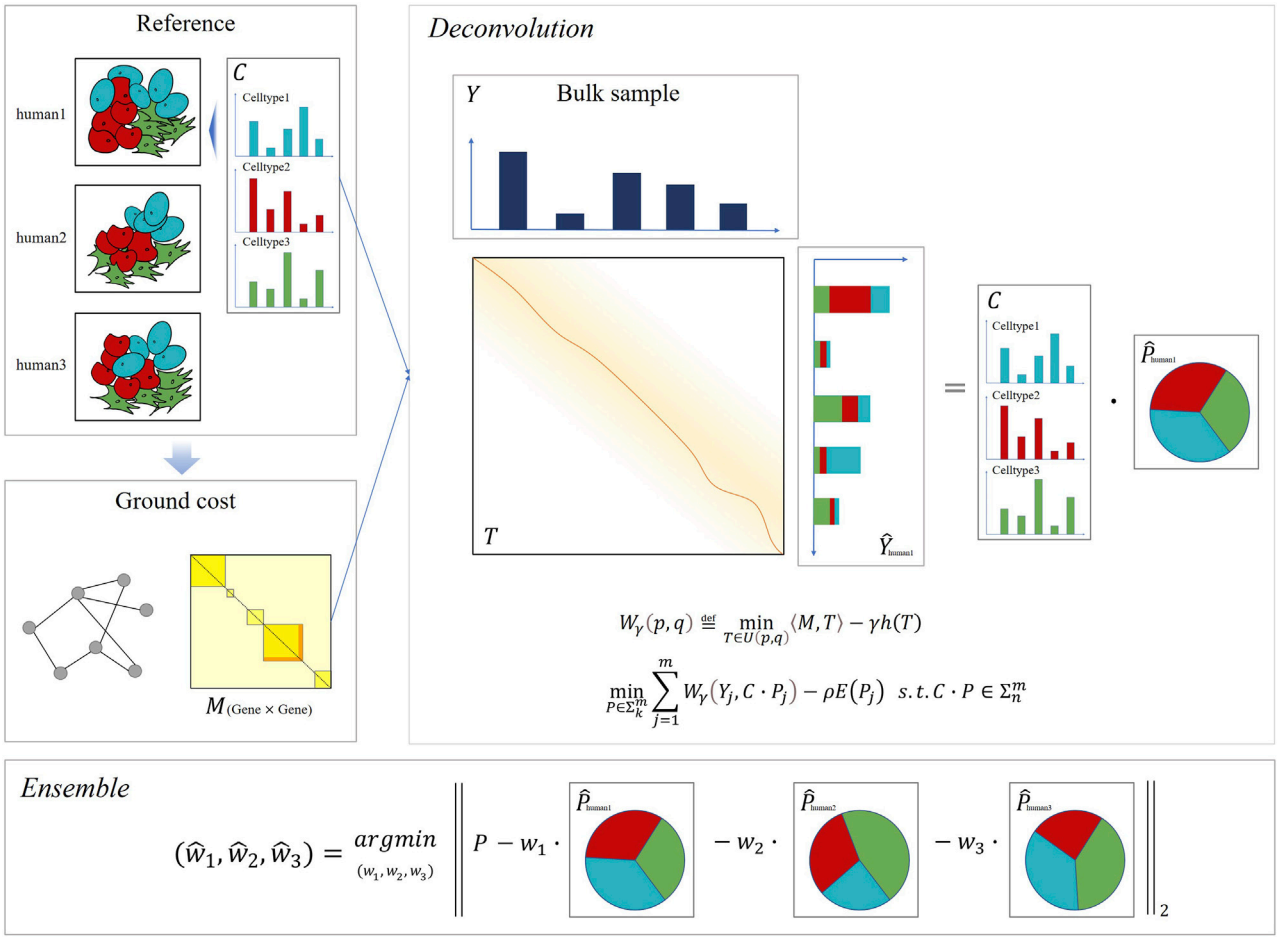


FIGURE 1 | Overview of DecOT. Based on the Wasserstein loss, DecOT first converts the gene expression of single-cell and bulk samples into distributions. Next, DecOT calculates the ground cost matrix from the single-cell expression data, which forms the correlations among genes, see *Ground Cost Selection* for details. Then, for each individual reference, DecOT finds the optimal proportion of cell types by minimizing the Wasserstein loss under the premise of a given ground cost. Finally, in order to resolve the impact of batch effects between individuals, DecOT uses an ensemble framework to weigh each individual's deconvolution result.

individual reference. In this way, the individual or batch effects can be accounted for simultaneously by DecOT.

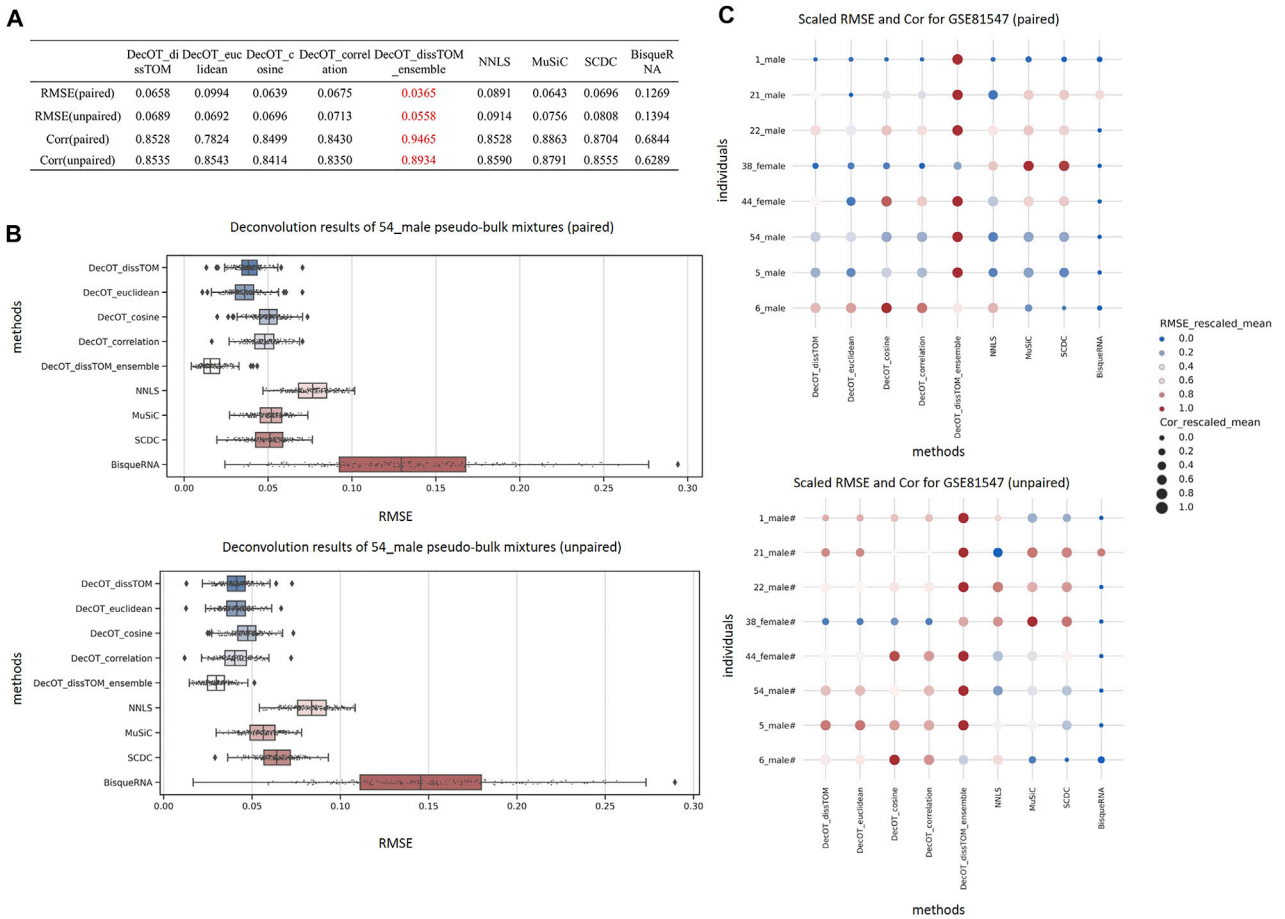
DecOT Outperforms Deconvolution Methods Based on Squared Loss

We evaluate DecOT with different ground costs as listed in *Ground Cost Selection*, which we refer to as DecOT_dissTOM, DecOT_euclidean, DecOT_cosine, and DecOT_correlation. For these four settings, we apply the aggregated reference, which is, pooling cells from multiple individuals to generate a single reference. In addition, we also evaluate DecOT with dissTOM under the ensemble framework (referred to as DecOT_disTOM_ensamble). The various settings of DecOT are then compared to four other square loss-based methods (including Nonnegative least squares (NNLS), MuSiC, SCDC, and Bisque) on artificial pseudo-bulk mixtures generated from four scRNA-seq data sets (**Table 1**, Methods). Since it is possible by design to assay both bulk-RNA and scRNA from the same individual (Kuksin et al., 2021), we consider settings of reference data in two situations:

a) There are annotated single-cell reference data from the same individual, from which the bulk sample is collected. We term such a situation as "paired".
b) Reference data are all collected from other individuals. We refer to such a scenario as "unpaired".

We mimic the "paired" situations in the benchmark by including cells (in the reference set) from the same individual for generating a pseudo-bulk sample (in the testing set) (**Supplementary Figure S1**).

Figure 2 shows the benchmark result of data set GSE81547 from Enge et al. (Enge et al., 2017) under these two situations. Applying DecOT under the ensemble framework has the best overall performance compared to other settings and methods. The average RMSE of DecOT_dissTOM_ensemble over all pseudo-bulks is 0.037 and 0.056 under paired and unpaired



	DecOT_dissTOM	DecOT_euclidean	DecOT_cosine	DecOT_correlation	DecOT_dissTOM_ensemble	NNLS	MuSiC	SCDC	BisqueRNA
RMSE(paired)	0.0658	0.0994	0.0639	0.0675	0.0365	0.0891	0.0643	0.0696	0.1269
RMSE(unpaired)	0.0689	0.0692	0.0696	0.0713	0.0558	0.0914	0.0756	0.0808	0.1394
Corr(paired)	0.8528	0.7824	0.8499	0.8430	0.9465	0.8528	0.8863	0.8704	0.6844
Corr(unpaired)	0.8535	0.8543	0.8414	0.8350	0.8934	0.8590	0.8791	0.8555	0.6289

FIGURE 2 | Benchmark results using data set GSE81547. **(A)** Average RMSE and Cor of the deconvolution results of all mixtures in data set GSE81547. From the overall results, the estimation of DecOT with the ensemble framework has smaller errors and stronger correlation than other methods. **(B)** Boxplot of RMSE from 200 replicate pseudo-bulk mixtures from sample 54_male. The top/bottom panel shows the results under paired/unpaired situation. **(C)** Overview of deconvolution results of individual pseudo-bulk mixtures across all methods in data set GSE81547. For each individual, we rank the results across different methods and rescale them to the interval between 0 and 1. A darker-red and larger point within a line represents a smaller RMSE and a larger Cor. Both paired (top) and unpaired (bottom) situations are considered.

situations, respectively, and the average correlation is 0.946 and 0.893 (**Figure 2A**). **Figure 2B** shows the detailed estimation results of individual sample 54_male in GSE81547. DecOT with an ensemble framework using dissTOM shows the greatest performance. Even when applying aggregated references, Wasserstein's loss still outperforms NNLS.

In order to show the overall quality of the various methods in pseudo-bulk mixtures generated from different samples in GSE81547, we compared the mean RMSEs and mean Cors, which result from performing different methods on the pseudo-bulk generated based on different individuals (**Figure 2C**). For each individual, we rank the results across different methods and rescale them to the interval between 0 and 1. As shown in **Figure 2C**, the dark-red and larger points within a line represent a smaller RMSE and a larger Cor. In general, DecOT using Wasserstein loss has better performance than square loss methods in most cases, and the ensemble framework can further improve the accuracy of the deconvolution results even when the mixtures and reference cells come from different individuals.

Similar conclusions are also obtained from benchmarks based on the other three data sets. The results are shown in **Supplementary Figures S2–S4**.

DecOT Performs Robustly Under the Ensemble Framework

The choice of reference in solving the supervised deconvolution problem is crucial. We first compare the performance of DecOT by using references from different individuals. In detail, we evaluate DecOT on the pseudo-bulk generated from the testing set of 54_male in GSE81547 by respectively applying reference data from each individual as well as under the ensemble framework (paired and unpaired). **Figure 3A** shows the result out of 200 pseudo-bulk mixtures in each reference setting. Using references from the same individual (reference set

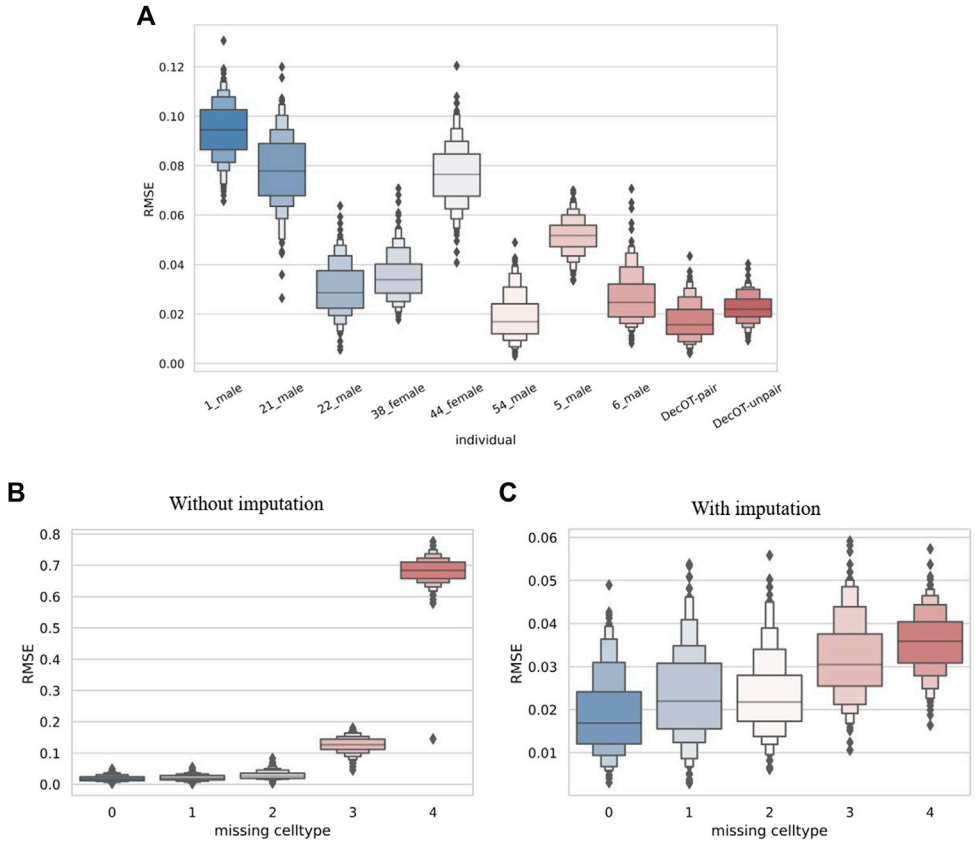


FIGURE 3 | Effects under different manipulation references benchmarked by mixtures constructed from 54_male of GSE81547. **(A)** Comparison of the results from single individual references and multi-individual references under the ensemble framework. **(B,C)** Deconvolution results with missing cell types in paired references. The cell types are progressively removed according to the ascending order of cell counts in 54_male. **(B)** Direct application of the paired reference from 54_male with the missing cell type. **(C)** Application of the paired reference from 54_male with missing cell types imputed by references from other individuals.

from 54_male) outperforms the situation of applying references from other individuals (**Figure 3A**). The deconvolution performance is slightly improved with integrating results across all individuals (paired), indicating that the DecOT ensemble framework makes use of information from other individuals to adjust the final estimation. Such a finding is further confirmed in the case under the unpaired reference situation; when excluding 54_male from the reference, the estimation of DecOT under the ensemble framework still obtains a smaller error than using other single individual references. In fact, including more individual references under the ensemble framework tends to improve the performance of deconvolution (**Supplementary Figure S5**).

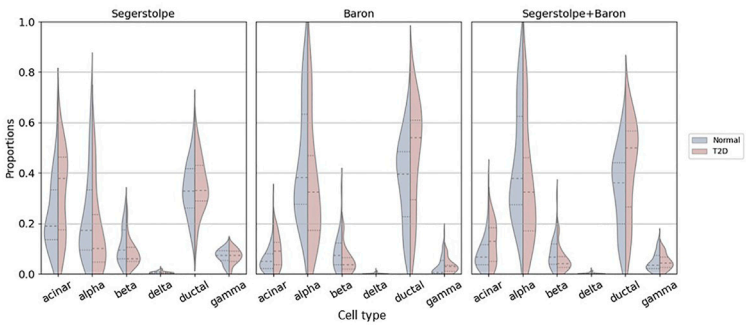
Deconvolution with paired single-cell data as a reference will greatly improve the performance. However, in a more realistic scenario, single cells collected from the same individual may have missing cell types as compared to the paired bulk sample, especially when the cell type is rare. Therefore, we conducted an experiment by gradually and cumulatively removing cell types in ascending order of cell count in the reference set of 54_male (**Supplementary Table S1**) and used the data with the missing cell type as a reference. When there is a missing cell type in the reference, the deconvolution may allocate the expression of the missing cell type to other types, which leads to biased estimation (**Figure 3B**). One way to reduce such bias is to impute the missing cell type in the reference by utilizing a publicly available data set as a surrogate. Here, we use the mean expression of the missing cell type from references of other individuals for imputation (**Figure 3C**). Compared to the results in **Figures 3B,C**, imputation of missing cell types significantly improves the performance of deconvolution. Nevertheless, regardless of imputation, the estimation error will get worse as the number of missing cell types increases.

Another possible way for reducing the impact caused by missing cell types in paired single-cell references is to apply DecOT under the ensemble framework. Since our ensemble framework integrates deconvolution results respectively performed under each individual reference, we can still apply imputation on missing cell types in the paired reference. **Table 2** compares the average RMSE of cases based on single references from paired single-cell data (RMSE-54_male) and ensemble references which account all possible individuals (RMSE-ensemble). In addition, we use the unpaired ensemble case as

TABLE 2 | Optimal weights of different individual references under the DecOT ensemble framework. The weights and the overall performance are compared under different settings of the missing cell type in the paired reference of sample 54_male. Imputation indicates that the reference profiles of missing types are imputed by references from other individuals.

	Optimal weight with imputation								RMSE-54_male	RMSE-ensemble
	1_male	21_male	22_male	38_female	44_female	54_male	5_male	6_male		
54_male-all	0.0000	0.0000	0.1654	0.0000	0.0000	0.7504	0.0842	0.0000	0.0190	0.0175
54_male-delta	0.0000	0.0000	0.1691	0.0000	0.0000	0.7412	0.0817	0.0081	0.0234	0.0215
54_male-delta-ductal	0.0000	0.0000	0.1772	0.0000	0.0000	0.7293	0.0830	0.0104	0.0234	0.0218
54_male-delta-ductal-acinar	0.0000	0.0000	0.1684	0.0000	0.0000	0.6985	0.0934	0.0397	0.0318	0.0289
54_male-delta-ductal-acinar-beta	0.0000	0.0000	0.1765	0.0000	0.0475	0.6251	0.1509	0.0000	0.0359	0.0306
54_male-unpair	0.0000	0.0000	0.5114	0.0000	0.1837	—	0.1487	0.1561	—	0.0227
	Optimal weight without imputation								**RMSE-54_male**	**RMSE-ensemble**
	1_male	**21_male**	**22_male**	**38_female**	**44_female**	**54_male**	**5_male**	**6_male**		
54_male-all	0.0000	0.0000	0.1462	0.0000	0.0000	0.7483	0.0859	0.0196	0.0190	0.0172
54_male-delta	0.0000	0.0000	0.1558	0.0000	0.0000	0.7310	0.0901	0.0231	0.0211	0.0165
54_male-delta-ductal	0.0000	0.0000	0.1625	0.0000	0.0000	0.7069	0.1093	0.0212	0.0293	0.0237
54_male-delta-ductal-acinar	0.0000	0.0000	0.2971	0.0000	0.0545	0.3624	0.2173	0.0686	0.1268	0.0444
54_male-delta-ductal-acinar-beta	0.0000	0.0000	0.4654	0.0000	0.1840	0.0081	0.1670	0.1755	0.6834	0.0226
54_male-unpair	0.0000	0.0000	0.4692	0.0000	0.1855	—	0.1683	0.1770	—	0.0224

A



B

	Segerstolpe		Baron		Segerstolpe+Baron	
	statistic	*p*-value	statistic	*p*-value	statistic	*p*-value
DecOT	-1.9108	0.0599	-1.9797	0.0514	-2.0107	0.0480
NNLS	0.3210	0.7491	-0.3981	0.6917	0.1010	0.9198
MuSiC	-1.3009	0.1973	-1.1173	0.2675	-1.7303	0.0877
SCDC	-1.3945	0.1673	-1.5831	0.1176	-1.8847	0.0633
BisqueRNA	1.0610	0.2921	1.0818	0.2828	1.1053	0.2725

C

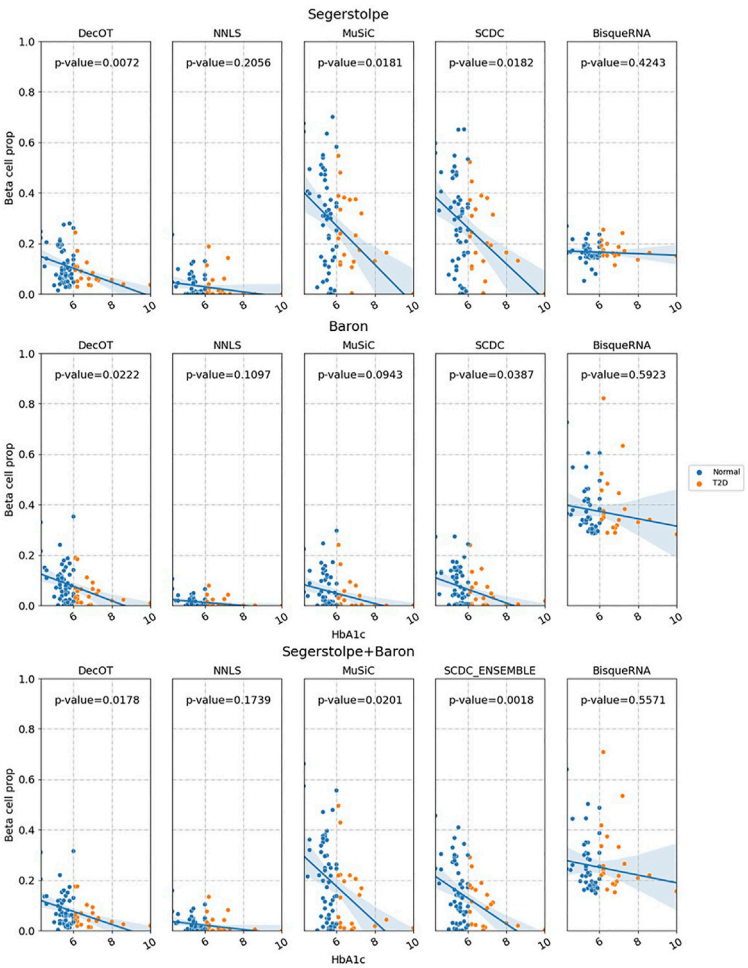


FIGURE 4 | Cell-type deconvolution of healthy and T2D human pancreatic islet samples. **(A)** Estimated composition of islet cell types in healthy and type 2 diabetes (T2D) humans by DecOT under three settings of references. The violin plots show the proportion differences between healthy and T2D samples. **(B)** Independent sample t-tests of beta cell proportion between healthy and T2D individuals. DecOT shows the most significant difference as compared to other methods. **(C)** Linear regression of HbA1c expression level and the proportion of beta cells estimated by five methods. The reported *p*-values come from a multivariate linear regression model: beta cell ratio ~ HbA1c + age + BMI + gender.

a baseline. The weight contributions of references from each individual are also displayed in **Table 2**. Since the pseudo-bulk mixtures are constructed from 54_male, the reference from one's own cell (self-ref) contributed the most to the ensemble result. The weight contribution from self-ref decreases with the increasing number of missing cell types. The ensemble DecOT estimation under the ensemble framework is always better than using a single reference, even though it is collected from the same individual as for the bulk sample. Such a result verifies that the ensemble framework can integrate the information of multiple individuals to get a better estimate even if there is a cell type missing in the paired reference. In general, the results from the ensemble framework are rather robust under missing cell types in paired references (regardless of whether they are imputed or not).

Performance of DecOT on Human Pancreatic Islet Data

Next, we apply DecOT with dissTOM as the ground cost to deconvolve the bulk samples of 89 human islets from Fadista et al. (Fadista et al., 2014), which contains 51 healthy individuals, 26 type 2 diabetic (T2D) individuals, and 12 unknown individuals. We focus on the composition of six cell types of interest (alpha, beta, delta, gamma, acinar, and ductal) in the human pancreatic islet. We use three groups of scRNA-seq references, denoted as the Baron reference (Avila Cobos et al., 2020), Segerstolpe reference (Segerstolpe et al., 2016), and ensemble reference, which combine data from both studies. **Figure 4A** shows the deconvolution results of DecOT on the six types of cells by contrasting the status of individuals (normal or T2D). The proportion of beta cells that secrete insulin will gradually decrease with the progression of type 2 diabetes (T2D) (Kanat et al., 2011), (Hou et al., 2015). DecOT can successfully detect such a proportion difference between normal and T2D patients, regardless of which group of reference is used for analysis. In addition, we also apply independent sample t-tests on the beta cell proportion estimated by DecOT between normal and T2D groups. The estimates of DecOT based on all three reference groups all result in significant differences in beta cell proportion between normal and T2D samples (**Figure 4B**). When comparing the results with those of the four other deconvolution methods, DecOT shows the most significant p-values (**Figure 4B**). Note that for the ensemble reference, SCDC applies its built-in ENSEMBLE method, which weighs the deconvolution results across two sources of references. The other methods directly use the pooled data as references.

Previous studies have shown that in human pancreatic islet samples, hemoglobin A1c (HbA1c) is an important biomarker of type 2 diabetes, and its expression level should be negatively correlated with beta cell functions (Kanat et al., 2011), (Hou et al., 2015), (Frogner et al., 2015). We perform linear regression to the estimates of beta cell proportion (BP) by HbA1c and adding age, gender, and BMI as covariates. **Figure 4C** shows the regression results. The estimates of BP by NNLS and BisqueRNA failed to recover a significant negative correlation to the level of HbA1c. The beta cell proportion estimated by DecOT, MuSiC, and SCDC based on the three groups of references discovered significant negative correlations with HbA1c. When using a single-source reference, DecOT calculated the smallest p-values (0.0599 and 0.0514), indicating a more significant correlation between BP and HbA1c levels. In fact, the estimated BP by DecOT is robust over all three groups of references, which can be seen from the variation between the slops of the fitted regression line in **Figure 4C**. In contrast, the slopes have greater differences in MuSiC and SCDC cases when a different reference is applied. In short, DecOT shows better performance on real data sets and is robust to different sources of references.

DISCUSSION

In this study, we proposed DecOT, which applies single-cell data as references and uses Wasserstein distance as a loss function for decomposing bulk cell types. Compared with the commonly used square loss methods, the optimization of Wasserstein loss in DecOT is able to utilize additional information from gene space, for example, ground cost induced by gene-gene relations. By benchmarking DecOT with four recently proposed square loss-based methods on pseudo-bulk data from four different single-cell data sets and real pancreatic islet bulk samples, DecOT shows superior performance.

Wasserstein loss accounts for the distance between genes through the ground cost matrix. In this study, we evaluated four possible choices of ground cost, namely, three common metrics (Euclidean distance, cosine similarity, and Pearson correlation) and the dissTOM distance based on gene co-expression networks. In the analysis of simulated data, the final deconvolution effect of the four metrics did not show much difference; however, since the topological overlap measure (TOM) has been considered a more robust measure of gene interconnections (Li and Horvath, 2007), we recommend using dissTOM over other metrics.

Although DecOT obtains better deconvolution accuracy by using Wasserstein loss, optimization of such a loss also brings a greater computational cost. The application of entropic regularization allows tractable computation of data sets on a larger scale. However, there is a trade-off between accuracy and computation time. This trade-off can be tuned by the two hyperparameters γ and ρ. In **Supplementary Figure S6**, we show the calculation time of DecOT under different numbers of genes and the accuracy and time of DecOT calculations under different choices of two regularization parameters. We show that the performance of DecOT is rather robust with parameters in the range of $\gamma \leq 0.05$ and $\rho \leq 0.01$, which results in higher calculation accuracy.

When applying a supervised bulk-cell-type deconvolution algorithm, the possible individual variation and batch effect should be noted when combining references from multiple individuals and/or data sets. DecOT uses an ensemble framework to weigh the deconvolution across multiple results from each individual reference to mitigate individual effects. The weights of the ensemble framework indicate, to a certain extent, the similarity of the gene distribution between the reference individuals and the bulk samples. In the benchmarks on pseudo-bulk data, DecOT using the ensemble framework shows improved accuracy and robustness over existing methods in most scenarios.

The performance of deconvolution will be greatly improved when paired single-cell references are available. However, there can be a problem regarding the cell-type integrity in the paired

reference. We have tested two solutions in the study, imputation of the missing cell types, and/or applying the ensemble framework with DecOT. The results show that the ensemble framework can effectively utilize information of missing cell types from other reference individuals by adjusting the weights. Although the imputation solution also achieves acceptable results, the ensemble framework of DecOT shows more robust performance.

DATA AVAILABILITY STATEMENT

All the data sets used for this study can be found at GitHub: https://github.com/lg-ustb/DecOT. These data sets are downloaded from their respective sources: GSE84133, GSE81547, E-MTAB-5061, GSE134355, and GSE50398 and https://figshare.com/articles/HCL_DGE_Data/7235471.

AUTHOR CONTRIBUTIONS

GL and LM designed research. GL, LM, and XL discussed the ideas and supervised the study. GL performed the research. GL and LM wrote the manuscript. All authors approved the final manuscript.

FUNDING

This work was supported by the National Key R&D Program of China (2019YFA0709501) and the National Natural Science Foundation of China (11971459; 31772435).

ACKNOWLEDGMENTS

We thank the editor and the reviewers for their helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.825896/full#supplementary-material

REFERENCES

Afshar, A., Yin, K., Yan, S., Qian, C., Ho, J. C., Park, H., et al. (2020). Swift: Scalable Wasserstein Factorization for Sparse Nonnegative Tensors. arXiv preprint arXiv:2010.04081.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein Generative Adversarial Networks," in International Conference on Machine Learning (PMLR), Sydney, Australia, 214–223.

Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data. *Nat. Commun.* 11 (1), 1–14. doi:10.1038/s41467-020-19015-1

Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K. (2018). Computational Deconvolution of Transcriptomics Data from Mixed Cell Populations. *Bioinformatics* 34 (11), 1969–1979. doi:10.1093/bioinformatics/bty019

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., et al. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cel Syst.* 3 (4), 346–360. doi:10.1016/j.cels.2016.08.011

Carithers, L. J., Moore, H. M., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2015). The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking* 13 (6), 307–308. doi:10.1089/bio.2015.29031.hmm

Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *Adv. Neural Inf. Process. Syst.* 26, 2292–2300.

Denisenko, E., Guo, B. B., Jones, M., Hou, R., de Kock, L., Lassmann, T., et al. (2020). Systematic Assessment of Tissue Dissociation and Storage Biases in Single-Cell and Single-Nucleus RNA-Seq Workflows. *Genome Biol.* 21 (1), 130. doi:10.1186/s13059-020-02048-6

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., et al. (2019). SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References. doi:10.1101/743591

Enge, M., Arda, H. E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K., et al. (2017). Single-cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* 171, 321–330. doi:10.1016/j.cell.2017.09.004

Fadista, J., Vikman, P., Laakso, E. O., Mollet, I. G., Esguerra, J. L., Taneera, J., et al. (2014). Global Genomic and Transcriptomic Analysis of Human Pancreatic Islets Reveals Novel Genes Influencing Glucose Metabolism. *Proc. Natl. Acad. Sci.* 111 (38), 13924–13929. doi:10.1073/pnas.1402665111

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., et al. (2021). Pot: Python Optimal Transport. *J. Machine Learn. Res.* 22 (78), 1–8.

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. (2015). Learning with a Wasserstein Loss. arXiv preprint arXiv:1506.05439.

Guo, G. (2020). HCL DGE Data.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., et al. (2020). Construction of a Human Cell Landscape at Single-Cell Level. *Nature* 581 (7808), 303–309. doi:10.1038/s41586-020-2157-4

Hou, X., Liu, J., Song, J., Wang, C., Liang, K., Sun, Y., et al. (2015). Relationship of Hemoglobin A1c withβCell Function and Insulin Resistance in Newly Diagnosed and Drug Naive Type 2 Diabetes Patients. *J. Diabetes Res.* 2016 (2015-11-10), 1–6. doi:10.1155/2016/8797316

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., et al. (2020). Accurate Estimation of Cell Composition in Bulk Expression through Robust Integration of Single-Cell Information. *Nat. Commun.* 11 (1). doi:10.1038/s41467-020-15816-6

Jin, H., and Liu, Z. (2021). A Benchmark for RNA-Seq Deconvolution Analysis under Dynamic Testing Environments. *Genome Biol.* 22 (1), 1–23. doi:10.1186/s13059-021-02290-6

Kanat, M., Winnier, D., Norton, L., Arar, N., Jenkinson, C., Defronzo, R. A., et al. (2011). The Relationship between β-Cell Function and Glycated Hemoglobin. *Diabetes Care* 34 (4), 1006–1010. doi:10.2337/dc10-1352

Kantorovich, L. V. (1942). On the Transfer of Masses. review of politics.

Kuksin, M., Morel, D., Aglave, M., Danlos, F.-X., Marabelle, A., Zinovyev, A., et al. (2021). Applications of Single-Cell and Bulk Rna Sequencing in Onco-Immunology. *Eur. J. Cancer* 149, 193–210. doi:10.1016/j.ejca.2021.03.005

Langfelder, P., and Horvath, S. (2008). Wgcna: an R Package for Weighted Correlation Network Analysis. *Bmc Bioinformatics* 9 (1), 559. doi:10.1186/1471-2105-9-559

Lee, D. D., and Seung, H. S. (1999). Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature* 401 (6755), 788–791. doi:10.1038/44565

Li, A., and Horvath, S. (2007). Network Neighborhood Analysis with the Multi-Node Topological Overlap Measure. *Bioinformatics* 23, 222–231. doi:10.1093/bioinformatics/btl581

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science* 297 (5586), 1551–1555. doi:10.1126/science.1073374

Rolet, A., Cuturi, M., and Peyré, G. (2016). "Fast Dictionary Learning with a Smoothed Wasserstein Loss" in Artificial Intelligence and Statistics (PMLR), Cadiz, Spain, 630–638.

Rolet, A., Seguy, V., Blondel, M., and Sawada, H. (2018). Blind Source Separation with Optimal Transport Non-negative Matrix Factorization. *EURASIP J. Adv. Signal. Process.* 2018. doi:10.1186/s13634-018-0576-2

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell RNA-Seq: Advances and Future Challenges. *Nucleic Acids Res.* 42 (14), 8845–8860. doi:10.1186/1755-8794-4-5410.1093/nar/gku555

Sandler, R., and Lindenbaum, M. (2011). Nonnegative Matrix Factorization with Earth Mover's Distance Metric for Image Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (8), 1590–1602. doi:10.1109/tpami.2011.18

Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B., and Raue, A. (2017). Estimation of Immune Cell Content in Tumour Tissue Using Single-Cell RNA-Seq Data. *Nat. Commun.* 8 (1), 1–12. doi:10.1038/s41467-017-02289-3

Schmitz, M. A., Heitz, M., Bonneel, N., Ngolè, F., Coeurjolly, D., Cuturi, M., et al. (2018). Wasserstein Dictionary Learning: Optimal Transport-Based Unsupervised Nonlinear Dictionary Learning. *SIAM J. Imaging Sci.* 11 (1), 643–678. doi:10.1137/17m1140431

Schmitz, M. A., Heitz, M., Bonneel, N., Ngolè, F., Coeurjolly, D., Cuturi, M., et al. (2018). Wasserstein Dictionary Learning: Optimal Transport-Based Unsupervised Nonlinear Dictionary Learning. *SIAM J. Imaging Sci.* 11 (1), 643–678. doi:10.1137/17M1140431

Segerstolpe, Å, Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., et al. (2016). Single-cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cel Metab.* 24, 593–607. doi:10.1016/j.cmet.2016.08.020

Sinkhorn, R. (1967). Diagonal Equivalence to Matrices with Prescribed Row and Column Sums. *The Am. Math. Monthly* 74 (4), 402–405. doi:10.2307/2314570

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). Review the Cancer Genome Atlas (TCGA): an Immeasurable Source of Knowledge. *wo* 1A (1A), 68–77. doi:10.5114/wo.2014.47136

Villani, C. (2009). *Optimal Transport: Old and New*, 338. Berlin: Springer, 23.

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference. *Nat. Commun.* 10 (1). doi:10.1038/s41467-018-08023-x

Weng, L. (2019). *From gan to Wgan*. arXiv preprint arXiv:1904.08994.

Yip, A. M., and Horvath, S. (2007). Gene Network Interconnectedness and the Generalized Topological Overlap Measure. *Bmc Bioinformatics* 8 (1), 22. doi:10.1186/1471-2105-8-22

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* 4 (1), Article17. doi:10.2202/1544-6115.1128

Zhang, S. Y. (2021). "A Unified Framework for Non-negative Matrix and Tensor Factorisations with a Smoothed Wasserstein Loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4203.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership