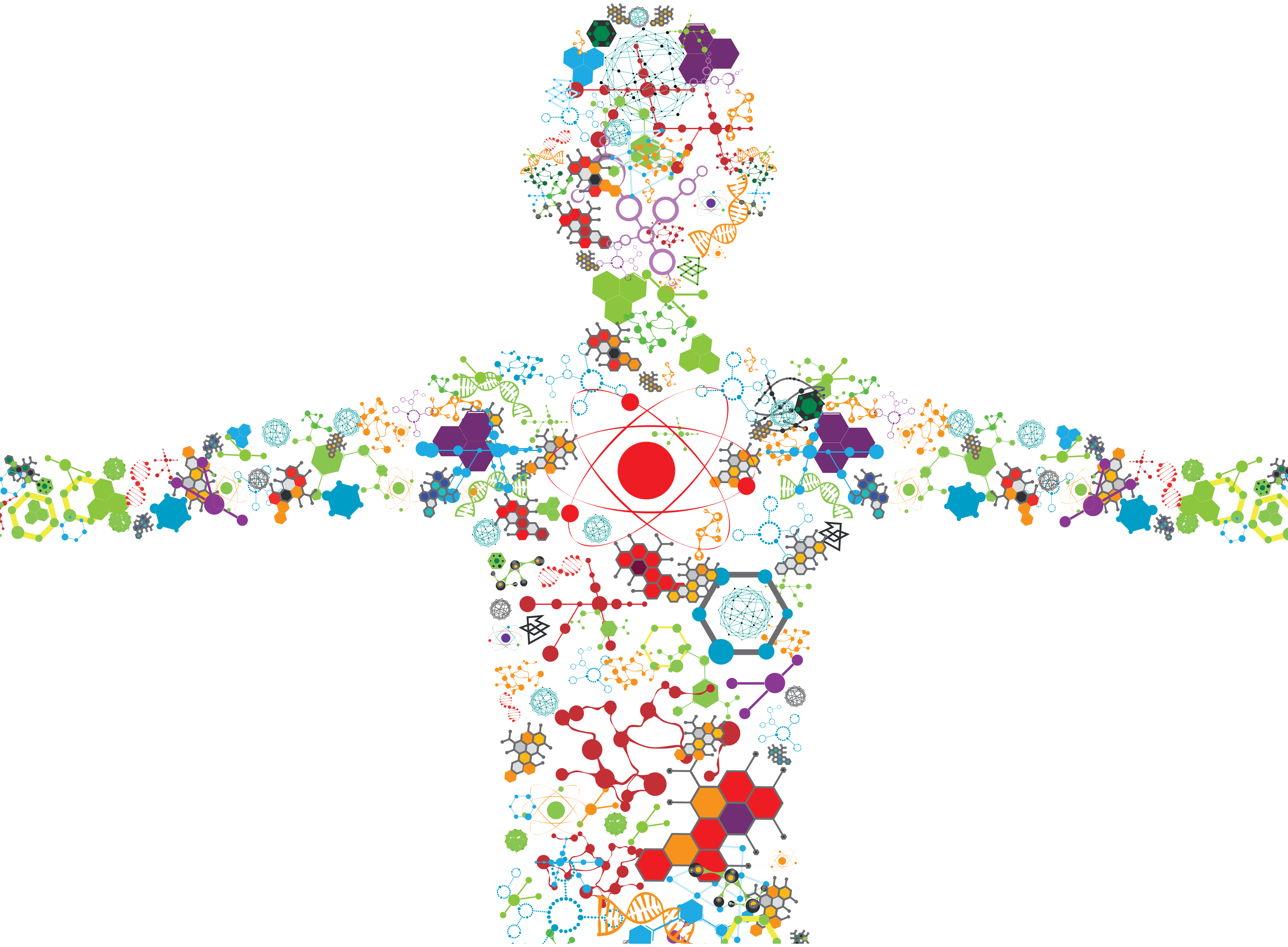
SINGLE CELL INTELLIGENCE AND TISSUE ENGINEERING

EDITED BY: Zhaoyuan Fang, Yangzi Jiang and Jiaofang Shao
PUBLISHED IN: Frontiers in Bioengineering and Biotechnology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-251-8
DOI 10.3389/978-2-83250-251-8

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

SINGLE CELL INTELLIGENCE AND TISSUE ENGINEERING

Topic Editors:
Zhaoyuan Fang, Zhejiang University, China
Yangzi Jiang, The Chinese University of Hong Kong, China
Jiaofang Shao, Nanjing Medical University, China

Citation: Fang, Z., Jiang, Y., Shao, J., eds. (2022). Single Cell Intelligence and Tissue Engineering. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-251-8

Table of Contents

TYPE Editorial
PUBLISHED 06 September 2022
DOI 10.3389/fbioe.2022.1019929

OPEN ACCESS

EDITED AND REVIEWED BY
Andrea Banfi,
University of Basel, Switzerland

*CORRESPONDENCE
Zhaoyuan Fang,
zhaoyuanfang@intl.zju.edu.cn

SPECIALTY SECTION
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

RECEIVED 15 August 2022
ACCEPTED 17 August 2022
PUBLISHED 06 September 2022

CITATION
Shao J, Jiang Y and Fang Z (2022), Editorial: Single cell intelligence and tissue engineering. *Front. Bioeng. Biotechnol.* 10:1019929. doi: 10.3389/fbioe.2022.1019929

Editorial: Single cell intelligence and tissue engineering

Jiaofang Shao[1], Yangzi Jiang[2,3,4,5] and Zhaoyuan Fang[6,7]*

[1]Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China, [2]Institute for Tissue Engineering and Regenerative Medicine, School of Biomedical Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China, [3]School of Biomedical Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China, [4]Department of Orthopaedics and Traumatology, Faculty of Medicine, The Chinese University of Hong Kong, and Prince of Wales Hospital, Shatin, Hong Kong SAR, China, [5]Center for Neuromusculoskeletal Restorative Medicine, Shatin, Hong Kong SAR, China, [6]Zhejiang University-University of Edinburgh Institute, Zhejiang University School of Medicine, Haining, China, [7]The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

KEYWORDS

single-cell, computational methods, biomarker identification, clinical application, profiling, machine learning

Editorial on the Research Topic
Single cell intelligence and tissue engineering

Single-cell sequencing has emerged as a powerful technology to dissect the heterogeneity of complex biological tissues at genomic, epigenomic, and transcriptomic levels, and has been extensively applied in various biological researches particularly in disease mechanisms and developmental biology (Paik et al., 2020; Gohil et al., 2021; Lei et al., 2021). Since the first single-cell RNA-sequencing (scRNA-seq) publication in 2009 (Tang et al., 2009), single-cell-based technologies have generated massive datasets, offering great opportunities to fully address biomedical problems as well as posing a challenge to computational analysis. At the same time, machine learning methods have been successfully used in processing many kinds of big data, including scRNA-seq data analysis (Petegrosso et al., 2020; Flores et al., 2022).

Nevertheless, more in-depth studies by using elegant methods and strategies to analyze the massive data obtained from the sequencing are still in need to improve our understanding of complex disorders. To make the best use of the single-cell-based data, researchers would first demand efficient and accurate computational pipelines to cluster, annotate cell types, uncover the marker genes and perform functional analysis. Besides, proper study design, including dataset selection and cross-validation, should be conducted to ensure that the evidence is convincing. In this context, this research topic included nine research articles focusing on methods development and clinical application of single-cell technology, giving more examples of data analysis and application in biomedical research.

One of the most common applications of single-cell approaches is to identify and distinguish cell types, and more related computational methods are demanded. Li et al. trained several classifiers and obtained optimal models from *in vitro* cultured human hepatocyte single-cell RNA data, and identified biomarkers for distinct differentiated hepatic cell types. By uncovering qualitative features for different stages of differentiation of liver cells, this study aimed to provide potential targets for cell transplantation to treat liver diseases. Similarly, Li et al. applied several different machine learning methods to expression profiling data of human pancreatic islet cells at single-cell resolution from both type 2 diabetes (T2D) patients and non-diabetic donors and discovered several T2D-associated genes. These two studies showed the promising applications of machine learning using single-cell expression profiling datasets to understand complex diseases.

Clustering is a critical step in single-cell data analysis to reveal heterogeneities and recognize cell types, requiring efficient and accurate computational algorithms. Tian et al. utilized an enhanced consensus-based clustering model and developed a novel computation method scMelody to cluster cells with single-cell DNA methylation data, such as scME-seq, scBS-seq, scWGBS, scTrio-seq, scNOMe-seq and snmC-seq. By using seven real single-cell methylation datasets and a variety of simulated datasets with different initial settings, scMelody showed better clustering performance and scalability when compared to other existing methods. In the two case studies, scMelody was able to uncover novel cell clusters from human hematopoietic cells and mouse neuron datasets.

Another two studies focused on developing prediction models for diagnosis using machine learning methods. Wang et al. explored mutation signatures with pan-cancer whole exome sequencing data, and constructed a logistic regression model to distinguish cancer types. The proposed model was able to trace the tumor origin for metastatic cancers, and predict cancer types using plasma ctDNA. Wu et al. investigated the microbiota in lung tissue and bronchoalveolar lavage fluid from lung ground-glass opacity (GGO) patients, and constructed a model using 10 genera-based biomarkers to predict GGO.

In addition to the biomarker identification for disease diagnosis, it has been widely recognized that the cellular heterogeneity is prevalent in either tissue or cultured cancer cells. Li et al. first identified a shared sub-cluster cancer stem cells (CSC) using 4 scRNA-seq datasets from upper gastrointestinal cancer (UGIC) patients including head and neck squamous cell carcinoma (HNSCC), esophageal cancer (EC), and gastric cancer (GC), and then compared the specific cells to scRNA-seq datasets from other 6 cancers including glioma, melanoma, osteosarcoma, breast cancer, ovarian cancer and stellate cell cancer to validate the specificity. The UGIC-specific CSC upregulated 33 genes while downregulated 141 genes compared to other tumors analyzed in this study, involving in inflammatory and Wnt pathways. Smit et al. applied FUNseq to spatially profile human breast epithelial cell line MCF10A which is a widely used *in vitro* model for breast cell transformation at single-cell resolution to decipher intratumor heterogeneity. By comparing the gene expression profiles and cell-cell communication among the cell populations at outer, middle and inner regions in 2D culture, the researchers found that cells at the outermost edge are most invasive, with epithelial-to-mesenchymal transition strongly activated.

While the advent of scRNA-seq technology has enabled profiling gene expression for each cell, analyses of large sample cohorts are still limited. The integration of existing bulk transcriptomic datasets is one of the issues that deserve attention and discussion in data mining. By taking advantage of the power of single-cell approaches and the sample size of bulk methods, two studies provided novel insights for diseases. Shi et al. collected a series of single-cell transcriptome datasets from non-failure hearts, dilated cardiomyopathy and ischemic cardiomyopathy hearts, followed by a comprehensive analysis to find out key genes involved in heart failure. Furthermore, the researchers obtained bulk gene expression datasets to validate the findings from single-cell datasets. On the other hand, Yao et al. developed a practical deconvolution pipeline by constructing a signature gene matrix which was then used to estimate cell proportion from bulk data with CIBERSORTx. Using preeclampsia microarray data, the researchers found that the proportion of trophoblast cells might contribute to the pathogenesis of the disorder.

As discussed above, articles in this special issue covered the fields of single-cell intelligence analysis methods development and computational model construction, providing more options for utilizing clinical datasets. By employing single-cell transcriptome data from patients, the researchers demonstrated the power of single-cell technology in important cell type recognition and key gene identification, broadening the clinical application of the technology. We envision that the use of advanced computational analysis approaches in single-cell datasets will reveal more useful and accurate biomarkers, and greatly benefit the diagnosis and treatments of complex diseases.

Author contributions

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

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.

References

Flores, M., Liu, Z., Zhang, T., Hasib, M. M., Chiu, Y. C., Ye, Z., et al. (2022). Deep learning tackles single-cell analysis-a survey of deep learning for scRNA-seq analysis. *Brief. Bioinform.* 23 (1), bbab531. doi:10.1093/bib/bbab531

Gohil, S. H., Iorgulescu, J. B., Braun, D. A., Keskin, D. B., and Livak, K. J. (2021). Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nat. Rev. Clin. Oncol.* 18 (4), 244–256. doi:10.1038/s41571-020-00449-x

Lei, Y., Tang, R., Xu, J., Wang, W., Zhang, B., Liu, J., et al. (2021). Applications of single-cell sequencing in cancer research: progress and perspectives. *J. Hematol. Oncol.* 14 (1), 91. doi:10.1186/s13045-021-01105-2

Paik, D. T., Cho, S., Tian, L., Chang, H. Y., and Wu, J. C. (2020). Single-cell RNA sequencing in cardiovascular development, disease and medicine. *Nat. Rev. Cardiol.* 17 (8), 457–473. doi:10.1038/s41569-020-0359-y

Petegrosso, R., Li, Z., and Kuang, R. (2020). Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief. Bioinform.* 21 (4), 1209–1223. doi:10.1093/bib/bbz063

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6 (5), 377–382. doi:10.1038/nmeth.1315

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 05 January 2022
doi: 10.3389/fbioe.2021.779225

Integrative Analysis of Bulk and Single-Cell RNA Sequencing Data Reveals Cell Types Involved in Heart Failure

Xin Shi[1†], Li Zhang[2†], Yi Li[1†], Jieyuan Xue[1], Feng Liang[1], Han-wen Ni[1], Xia Wang[1], Zhaohua Cai[1], Ling-hong Shen[1]*, Tao Huang[3]* and Ben He[1]*

[1]Department of Cardiology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China, [2]Key Laboratory of Advanced Theory and Application in Statistics and Data Science, East China Normal University, Ministry of Education, Shanghai, China, [3]Bio-Med Big Data Center, Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Yangzi Jiang,
The Chinese University of Hong Kong,
China

Reviewed by:
Shining Ma,
Stanford University, United States
Michele Carrabba,
University of Bristol, United Kingdom

***Correspondence:**
Ling-hong Shen
rjshenlinghong@126.com
Tao Huang
huangtao@sibs.ac.cn
Ben He
drheben@126.com

†These authors share first authorship

Specialty section:
This article was submitted to
"Preclinical Cell and Gene Therapy",
a section of the journal
Frontiers in Bioengineering and
Biotechnology

Received: 18 September 2021
Accepted: 03 December 2021
Published: 05 January 2022

Citation:
Shi X, Zhang L, Li Y, Xue J, Liang F, Ni H-w, Wang X, Cai Z, Shen L-h, Huang T and He B (2022) Integrative Analysis of Bulk and Single-Cell RNA Sequencing Data Reveals Cell Types Involved in Heart Failure.
Front. Bioeng. Biotechnol. 9:779225.
doi: 10.3389/fbioe.2021.779225

Owing to the high mortality rates of heart failure (HF), a more detailed description of the HF becomes extremely urgent. Since the pathogenesis of HF remain elusive, a thorough identification of the genetic factors will provide novel insights into the molecular basis of this cardiac dysfunction. In our research, we performed publicly available transcriptome profiling datasets, including non-failure (NF), dilated cardiomyopathy (DCM) and ischemic cardiomyopathy (ICM) hearts tissues. Through principal component analysis (PCA), gene differential expression analysis, gene set enrichment analysis (GSEA), and gene Set Variation Analysis (GSVA), we figured out the candidate genes noticeably altered in HF, the specific biomarkers of endothelial cell (EC) and cardiac fibrosis, then validated the differences of the inflammation-related cell adhesion molecules (CAMs), extracellular matrix (ECM) genes, and immune responses. Taken together, our results suggested the EC and fibroblast could be activated in response to HF. DCM and ICM had both commonality and specificity in the pathogenesis of HF. Higher inflammation in ICM might related to autocrine CCL3/CCL4-CCR5 interaction induced chemokine signaling activation. Furthermore, the activities of neutrophil and macrophage were higher in ICM than DCM. These findings identified features of the landscape of previously underestimated cellular, transcriptomic heterogeneity between ICM and DCM.

Keywords: single-cell RNA sequencing, transcriptome, heart failure, dilated cardiomyopathy, ischemic cardiomyopathy

INTRODUCTION

Heart failure (HF) is a chronic, progressive syndrome with high mortality and mobility, and affects approximately over 37.7 million patients worldwide (Ziaeian and Fonarow, 2016). HF is a serious process of cardiac dysfunction, characterized by impairment of ejection of blood or ventricular filling or both. HF brings a considerable burden to the health-care system, and leads to high rates of hospitalizations, readmissions, and outpatient visits (Bui, Horwich, and Fonarow, 2011; Jones, Roalfe, Adoki, Hobbs, and Taylor, 2019). The rising incidence of HF is associated with multiple factors (Triposkiadis, Xanthopoulos, and Butler, 2019), including age, obesity, hypertension, diabetes mellitus, ischemic heart disease, comorbidities, heredity, and environment, making it difficult to blame it on one specific issue (Oneglia, Nelson, and Merz, 2020; Triposkiadis, Xanthopoulos, Parissis, Butler, and Farmakis, 2020). Since HF is associated with high and unpredictable mortality, there is an emerging interest in potential HF biomarkers, and this exploration benefits the strategies of scientific prevention and advanced therapy.

Complex biological processes are involved in the pathogenesis of HF, and cardiac abnormalities often lead to heart dysfunction. Liu et al.(Liu et al., 2015) collected and analyzed left ventricle issues from six individuals including one ISCH patient, two dilated cardiomyopathy (DCM) patients and three controls as training sets to reveal genetic signatures of HF using RNA-seq and microarray data, which were further validated by a larger cohort with 313 individuals with HF or non-failing (NF). (Sweet et al., 2018) utilized RNA-seq and pathway analysis to reveal the heterogeneous gene signatures and disease-specific mechanisms in 64 explanted human hearts, which consisted of 37 DCM patients, 13 ICM patients, and 14 NF controls. (Vigil-Garcia et al., 2020) applied cardiomyocyte-specific transcriptomic analysis to detect a specific gene set involved in the process of pathological cardiac remodeling related to HF, and they explained the alternations precisely, which occurred during the transition from hypertrophic towards failing cardiomyocytes.

The advances in single-cell RNA sequencing (scRNA-seq) technology offers us an alternative method to characterize cell types involved in HF at the molecular level, which enables its broad application in HF research. (Yamaguchi et al., 2020) manifested that D1R signaling played a pathogenic effect on the process of HF, and explained the association between the activation of D1R and increased risk of patients with HF, using a mouse model of pressure overload-induced HF and single-cell resolution analysis, which aimed to uncover gene expression changes in murine models and human patients at the early and the late stages of HF. (Martini et al., 2019). used single-cell RNA sequencing data to describe the cardiac immune microenvironment in the heart of mouse models with the pressure-overload transverse aortic constriction (TAC) at early and late time points, providing novel diagnostic or therapeutic targets strategies for HF. However, as the sample size of scRNA-seq data is relatively small, and the mechanistic investigation in the variations of some cell types and cell type specific genes involved in HF required the integrative analysis of scRNA-seq and bulk RNA-seq data. In this study, we tried to identify some novel cell types, cell type specific genes and key components in HF by integrating bulk and single-cell RNA sequencing data, and anticipated to reveal cell types involved in DCM and ICM, which will offer a clearer demonstration of the immune inflammation response of HF.

MATERIALS AND METHODS

Data Collection

The single-cell RNA-seq data of two normal left ventricle samples were collected from Gene Expression Omnibus (GEO) with accession number GSE134355 (Han et al., 2020). To identify cell types and key genes related to heart failure, we downloaded the single-cell RNA-seq data of two normal, four dilated cardiomyopathy (DCM), and two ischemic cardiomyopathy (ICM) hearts samples (accession number: GSE121893 (Wang et al., 2020)), one scRNA-seq data of one normal, two DCM and two ICM hearts (accession number: GSE145154 (Rao et al., 2021)) for validation, and bulk RNA-seq data of 14 non-failure (NF), 37 DCM, and 13 ICM samples from GEO database (accession number: GSE116250 (Sweet et al., 2018)). The RNA-seq data of fibroblasts induced by TGFβ1 and control samples, and the microarray-based gene expression data for validation were downloaded from GEO with accession numbers GSE97358 (Schafer et al., 2017) and GSE5406 (Hannenhalli et al., 2006), respectively.

Cell Clustering Analysis

The unique molecular identifiers (UMIs) count-based scRNA-seq data of the two normal left ventricle samples were used for the cell clustering analysis, which was implemented in R Seurat v3.2.3 package. Cells with less than 500 UMIs were eliminated and features detected in less than 3 cells were filtered. The two hearts were integrated using the anchors by Reciprocal PCA. The expression data was normalized by LogNormalize method with scale factor = 1000,000, and top 2000 highly variable features were selected by FindVariableFeatures with dispersion method. The clusters were found at a resolution of four by FindClusters, and T-distributed Stochastic Neighbor Embedding (t-SNE) was applied to reduce the dimensionality. The cell-type marker genes were detected by FindAllMarkers function at adjusted p-value < 0.05, minimal percentage >0.25, and log2 fold change >0.25. All the marker genes of the cell clusters were collected from the earlier study (Han et al., 2020). This analysis was implemented by R Seurat v3.2.3 package (Stuart et al., 2019).

Principal Component Analysis for the Bulk RNA-Seq Data

The bulk RNA-seq data was downloaded from GEO database (GEO accession number: GSE116250 (Sweet et al., 2018)). The FPKM-based gene expression data were used for PCA analysis. Specifically, gene expressions higher than 1 FPKM in more than five samples were transformed to log2 (FPKM + 1), and the principal components were calculated by R FactoMineR package (Le, Josse, and Husson, 2008) and visualized by R factoextra package.

Gene Differential Expression Analysis

The pre-normalized microarray data and the RNA-seq data normalized to log2 (FPKM or RPKM +1) were tested by student t test and fold change. The count-based RNA-seq data was processed in R/Bioconductor DESeq2 package (Love, Huber, and Anders, 2014). All p-values were adjusted using the Benjamini and Hochberg approach. Genes with an adjusted p-value less than 0.05 and a fold change more than two were deemed as differentially expressed genes. Those genes could be ranked by the student t statistic to measure the differential expression levels.

Identification of Cell-types Involved in Heart Failure

The upregulated or downregulated genes in DCM/ICM samples were used for the identification of cell types significantly altered in

HF. The gene set overrepresentation enrichment analysis (Fisher's exact test) was employed to evaluate the significance of the differentially expressed genes (DEGs) against the cell type specific marker genes, which was implemented in R clusterProfiler (Yu, Wang, Han, and He, 2012) package.

Identification of Endothelial Cell Specific Marker Genes and Cardiac Fibrosis-Related Genes in HF

The gene set enrichment analysis (GSEA) was used to calculate the enrichment degree of those upregulated genes involved in HF or cardiac fibrosis in endothelial cells. Specifically, all the genes were pre-ranked by the t statistic, which represented the differential expression levels. The GSEA analysis was implemented in R clusterProfiler (Yu et al., 2012), and the genes identified as core enrichment in this analysis were considered as key components.

Gene Set Enrichment Analysis

The gene set overrepresentation enrichment analysis (ORA) was employed to identify the Reactome pathways enriched by previously detected endothelial cell specific marker genes and cardiac fibrosis-related genes in HF. This analysis was implemented in R ReactomePA package and visualized by R clusterProfiler (Yu et al., 2012) package.

The Cell Activity Estimation

The cell activity was estimated using single-sample Gene Set Variation Analysis (Hanzelmann, Castelo, and Guinney, 2013) (GSVA). Specifically, gene expression profiles and cell type specific marker genes were used as the input for GSVA to estimate the relative activities for each cell type and each sample.

Statistical Analyses

The two-sample comparison was conducted by student *t* test, and the multiple-sample comparison was implemented by analysis of variance (ANOVA). The *p*-values for multiple-sample comparisons were adjusted to *q*-values by the Benjamini and Hochberg method. Any *p*-values or *q*-values less than 0.05 were considered as statistically significant.

RESULTS

Identification and Characterization of Cell Types in Human Left Ventricle

To identify and characterize the cell types in the human left ventricle (LV), we collected two single-cell RNA sequencing datasets (scRNA-seq) of left ventricle provided by earlier study (Han et al., 2020). Subsequently, we eliminated the cells with low quality and retained 1,324 and 1,480 cells for further analysis (Materials and methods). As shown in **Figure 1A**, the cells from the two hearts were clustered into 18 clusters by the T-distributed Stochastic Neighbor Embedding (t-SNE) analysis, respectively. Using scHCL method, we successfully annotated 11 cell types for the two hearts (**Figure 1A**). Notably, the marker genes were specifically expressed in the cell types (**Figure 1B**). These results indicated that the cell types in the human left ventricle tissues could be identified and well-characterized by the scRNA-seq data.

The Cell Type Marker Genes Significantly Altered in Heart Failure

With the cell types and marker genes in the left ventricles, we aimed to identify the cell types altered in the left ventricles of heart failure. We analyzed the gene expression profiles of 14 NF, 37 DCM, and 13 ICM samples from previous study (Sweet et al., 2018). The PCA and differential expression analysis revealed that the samples from the three groups exhibited significantly different expression patterns (**Figures 2A,B**). Furthermore, we also conducted GSEA on the marker genes of cell types to test whether those marker genes were clustered within the upregulated or downregulated genes of ICM or DCM. Specifically, the marker genes of fibroblast and endothelial cell were significantly enriched within the upregulated genes in both DCM and ICM (**Figure 2C**, adjusted *p*-value < 0.05), suggesting that the dysfunction of the two cell types might be associated with both DCM and ICM. Moreover, marker genes of dendritic cell, M1/2 macrophage, neutrophil, and smooth muscle cell were more specifically enriched within the upregulated genes in ICM (**Figure 2C**, adjusted *p*-value < 0.05). These results indicated that DCM and ICM had both similarity and specificity in the pathogenesis of heart failure based on these disease-related cell types.

Key Regulators in the Endothelial Cells and Fibroblasts of Heart Failure

As the endothelial cell and fibroblast could be activated in response to HF (Colombo et al., 2005), we then investigated the key regulators in the ECs and fibroblasts of HF, and collected scRNA-seq data of 1,082 endothelial cells from the left ventricles of NF, DCM, and ICM samples (Wang et al., 2020). The comparison of DCM and ICM samples with NF samples revealed that the endothelial cell specific marker genes were highly enriched in the upregulated genes of HF endothelial cells (**Figure 3A**, FDR <0.05). Specifically, a total of 24 EC marker genes were found to be upregulated in both HF tissues (bulk RNA-seq) and the endothelial cells of HF samples (scRNA-seq) (**Figure 3B**, *p*-value < 0.05). The pathway enrichment analysis identified inflammation-related cell adhesion molecules (CAMs) as key regulators, including *CD74, HLA-B, HLA-E, HLA-DRB1, HLA-DQA1, HES1* and *CLDN5*, involved in the pathogenesis of HF (**Figure 3C**, FDR <0.05).

Furthermore, as transforming growth factor β1 (TGFβ1) is the principal pro-fibrotic factor in fibroblast activation (Akhurst & Hata, 2012), (Davis & Molkentin, 2014), which played vital roles in cardiac fibrosis (Ma, Iyer, Jung, Czubryt, & Lindsey, 2017), we examined whether the upregulated fibroblast marker genes in HF were involved in cardiac fibrosis. Consistently, we identified a large proportion of fibroblast marker genes upregulated in TGFβ1 induced cardiac fibroblast by differential expression analysis and GSEA (**Figure 4A**, FDR <0.05). Among these fibroblast marker

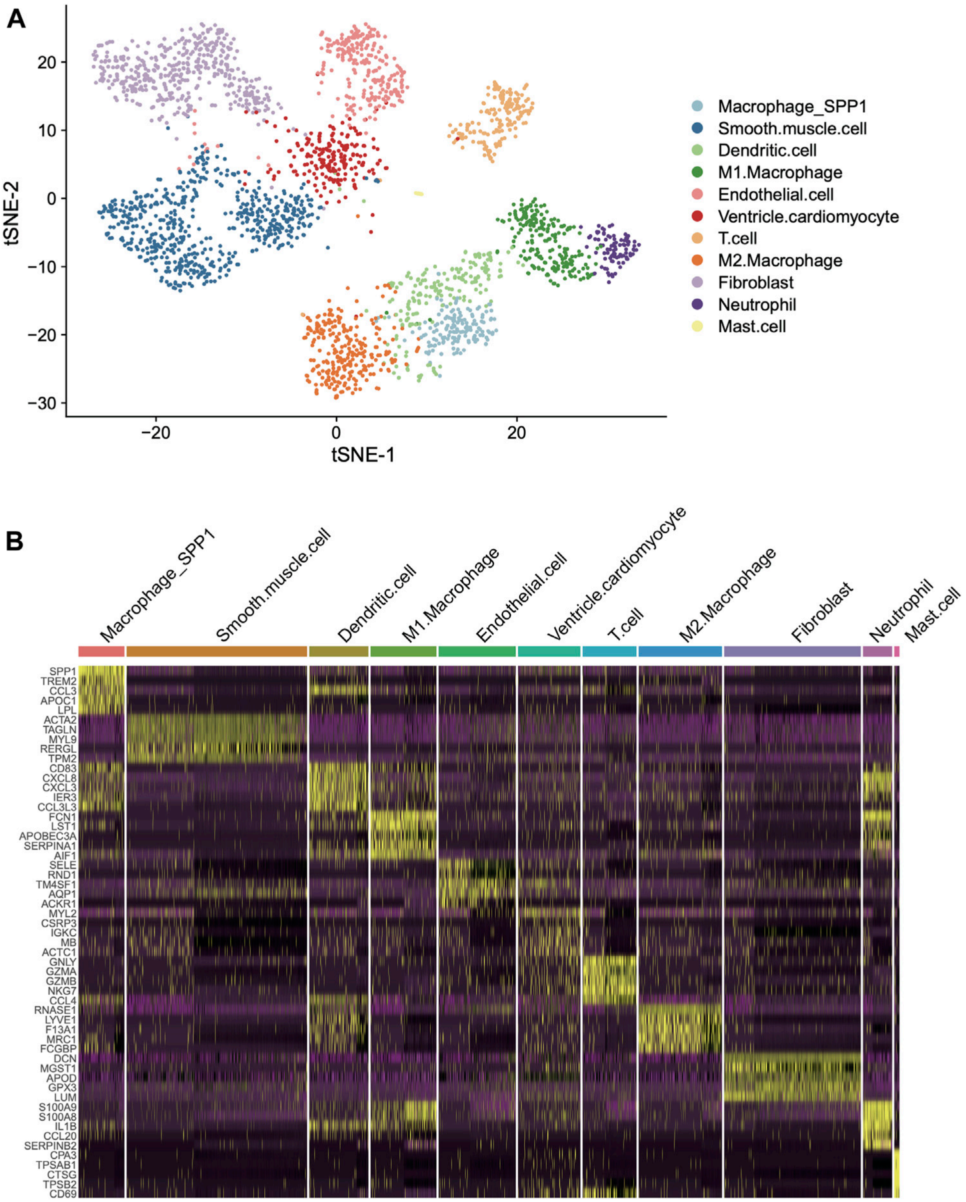


FIGURE 1 | Classification and molecular characterization of the cell types in two human left ventricles. **(A)** The T-distributed Stochastic Neighbor Embedding (t-SNE) analysis for the two left ventricles. Each point represents one cell, and the point colors represent the cell types. **(B)** The expression patterns of the cell type specific maker genes across the cell types in the two hearts (left ventricles).

genes, 29 were also upregulated in both HF tissues and fibroblast with TGFβ1 treatment (**Figure 4B**, FDR <0.05). The functional characterization of these genes revealed that *LTBP2*, *LTBP1*, *COL3A1*, *MFAP4*, *COL12A1*, *COL1A1*, *COL1A2*, *MMP2*, *TIMP2*, and *PCOLCE2* were primarily involved in extracellular matrix (ECM) organization and collagen biogenesis/formation/degradation (**Figure 4C**, FDR <0.05). Collectively, these results indicated that inflammation-related CAMs and ECM proteins such as collagens were specifically secreted by endothelial cell and fibroblast, respectively, and might induce cardiac inflammation and fibrosis during heart failure.

Chemokine Signaling Activation is Associated with Higher Inflammation in ICM

As ICM had more specific immune cell types, such as macrophage and dendritic cell (DC), than DCM, we then estimated the activities of immune cells including macrophage, DC, and neutrophil. Neutrophil and macrophage appeared to have higher activities in ICM than DCM and NF (**Figure 5A**, *p*-value < 0.05). Consistently, the marker genes of neutrophil and macrophage were also observed to be specifically upregulated in ICM (**Figure 5B**, *p*-value < 0.05). The cell-cell communication

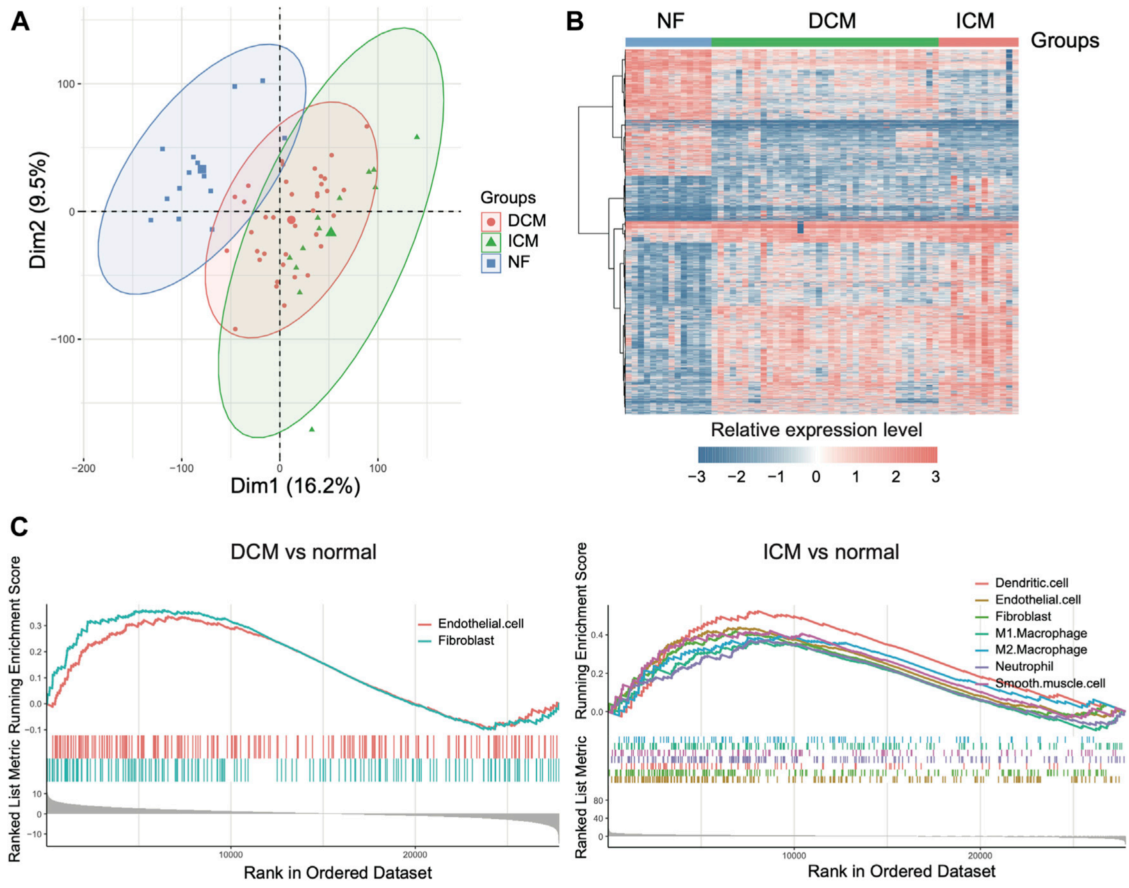


FIGURE 2 | The differentially expressed genes in dilated cardiomyopathy (DCM) and ischemic cardiomyopathy (ICM). **(A)** The scatterplot of principal component analysis for the samples. **(B)** The expression profiles of the differentially expressed genes (DEGs) in DCM and ICM. **(C)** The marker genes of cell types enriched within the upregulated genes of DCM or ICM.

analysis revealed that the autocrine ligand-receptor interaction induced chemokine signaling activation in neutrophil and macrophage might be responsible for the immune response in ICM (**Figure 5C**). Particularly, the ligands, CCL3, and CCL4, and the receptor CCR5 were specifically upregulated in ICM as compared with DCM and normal controls (**Figure 5D**). These results indicated that higher inflammation in ICM might be associated with autocrine CCL3/CCL4–CCR5 interaction induced chemokine signaling activation.

Validation of the Inflammation-Related CAMs, ECM Genes, and Immune Responses in an Independent Dataset

We collected an independent gene expression dataset from previous study (Hannenhalli et al., 2006) for validation. The inflammation-related CAMs such as HLA-E, HLA−DQA1, HLA−DRB1, and CD74, and all the ECM genes were upregulated in the HF samples of bulk RNA-seq dataset (GSE121893, **Figure 6A**, p-value < 0.05). Notably, the ECM genes were also upregulated in the fibroblasts of HF from an independent scRNA-seq dataset (**Figure 6B**). Furthermore, neutrophil and macrophage activities also appeared to be higher in ICM compared with NF and DCM, and the upregulation of autocrine ligand-receptor pairs in ICM, CCL3/CCL4 –CCR5, was also observed in the validation dataset (**Figures 6C,D**, p-value < 0.05). Consistently, the CCL3 and CCL4 were expressed higher in the macrophages of ICM than the DCM and normal hearts (**Figure 6E**). These results further indicated that inflammation-related CAMs and ECM proteins, which were specifically secreted by endothelial cell and fibroblast, respectively, and chemokine signaling activation in neutrophil and macrophage might induce cardiac inflammation and fibrosis during heart failure.

DISCUSSION

HF is a major consequence of various cardiovascular diseases with poor prognosis and high mortality (Shantsila, Wrigley, Blann, Gill, & Lip, 2012). In the present study, in order to clarify the cell heterogeneity between ischemic HF and non-ischemic HF, we integrated two scRNA-seq datasets of 1,324 and 1,480 cells from the left ventricles and gene expression profiles of 14 NF, 37 DCM, and 13 ICM samples to identify HF-related cell types and key

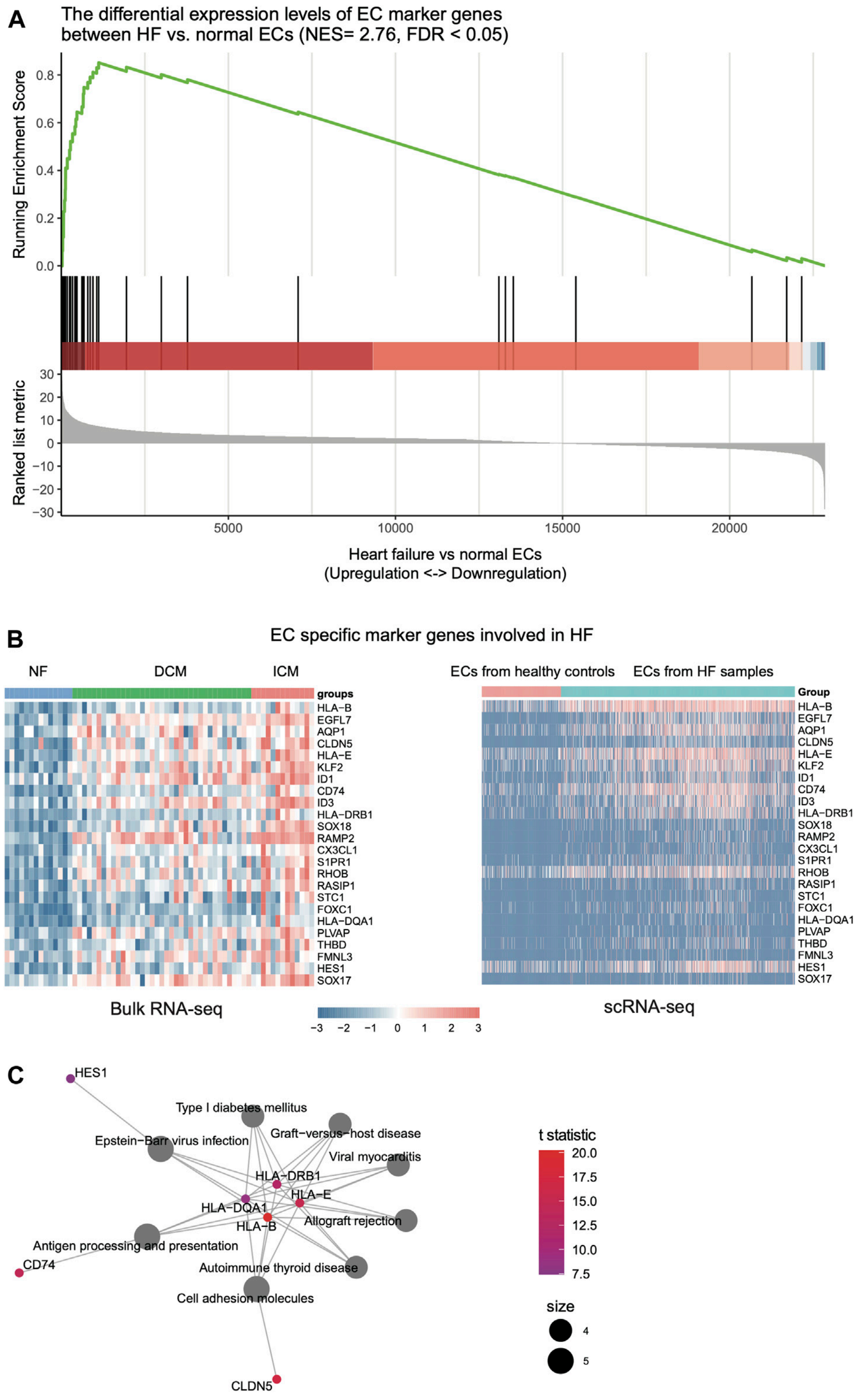


FIGURE 3 | The expression patterns of endothelial cell (EC)-related key regulators involved in HF. **(A)** The genes specifically upregulated in ECs of HF, which are identified by the gene set enrichment analysis (GSEA). **(B)** The expression patterns of genes in bulk RNA-seq and scRNA-seq data of ECs. **(C)** The key regulators in ECs by gene set enrichment analysis (GSEA).

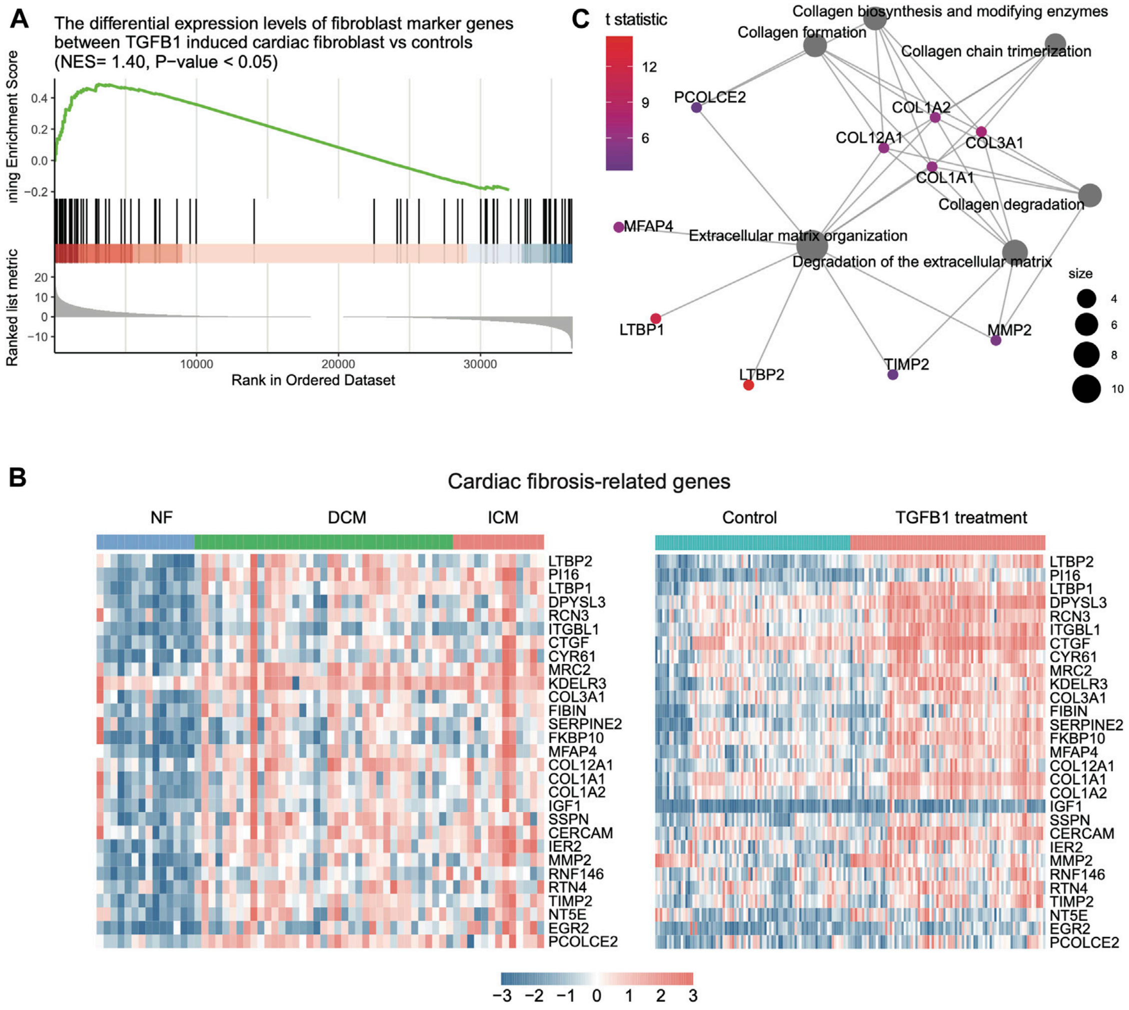


FIGURE 4 | The expression patterns of fibroblast-related key regulators involved in HF. **(A)** The genes specifically upregulated in TGF-beta-induced fibroblast by gene set enrichment analysis (GSEA). **(B)** The expression patterns of cardiac fibrosis-related genes in bulk RNA-seq and scRNA-seq data. **(C)** The key regulators involved in cardiac fibrosis by gene set enrichment analysis (GSEA).

regulators. Specifically, the marker genes of ECs were significantly upregulated in DCM and ICM proposing that the endothelial dysfunction might be associated with both DCM and ICM. In contrast, DC, M1/2 macrophage, neutrophil, and smooth muscle cell, were specifically upregulated in ICM based on the biomarkers of cell subpopulations. ECs are the most abundant non-myocytes in the healthy heart (Bacmeister et al., 2019). The patterns of endothelial dysfunction in HF patients differed from the etiologies (Oatmen, Cull, and Spinale, 2020). In patients with ischemic HF, endothelial dysfunction is systemic and involves both arteries and veins, conductance vessels and microvascular beds, coronary, pulmonary, and peripheral vessels, however, the patterns of endothelial dysfunction in non-ischemic HF are heterogeneous with fewer features of systemic abnormalities which have a functionally preserved endothelium in peripheral arteries (Berezin, Kremzer, Martovitskaya, Berezina, & Gromenko, 2016).

Fibroblasts as the main effector cells of cardiac fibrosis will be activated after injury associated with HF and participate the process of repair and remodel the infarcted heart (Davis & Molkentin, 2014). Cardiac fibrosis is characterized by an increased amount and a disrupted composition of inflammation-related CAMs and ECM proteins which might be potential targets for heart repair and function (Humeres & Frangogiannis, 2019; Moore-Morris, Guimaraes-Camboa, Yutzey, Puceat, & Evans, 2015). TGF-β1 as a cytokine could induce the transformation of cardiac fibroblasts to myofibroblasts (Akhurst & Hata, 2012). We examined whether the upregulated fibroblast marker genes in HF were involved in cardiac fibrosis through GSEA and differential expression analysis. Among these fibroblast marker genes, 29 were also upregulated in both HF tissues and fibroblast with TGFβ1 treatment. The functional characterization of these genes revealed that they were primarily involved in

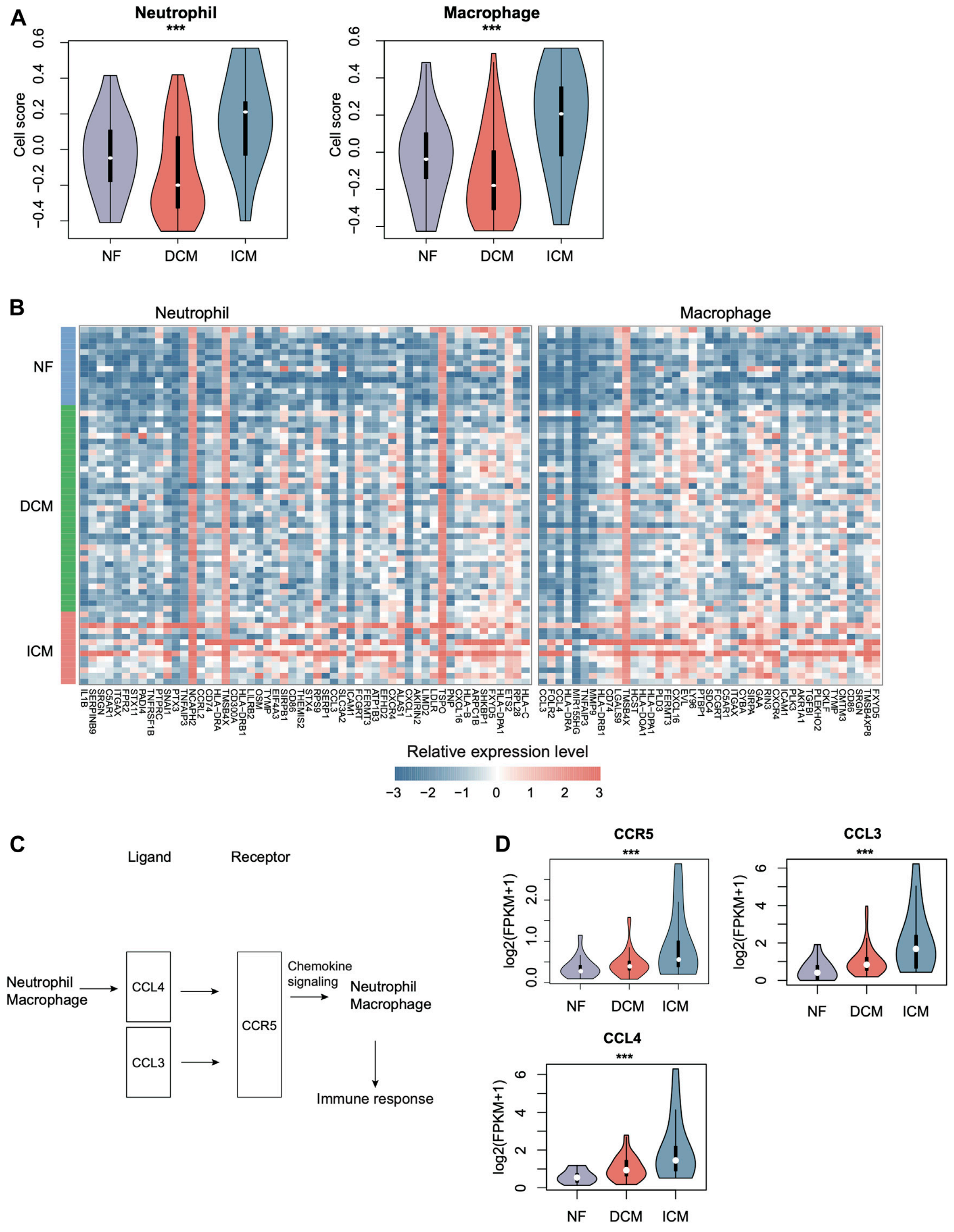


FIGURE 5 | The specific expression patterns of immune cell marker genes in ICM. **(A)** The relative abundances of immune cells including neutrophil and macrophage across the groups. **(B)** The expression patterns of immune cell-specific marker genes in NF, DCM, and ICM samples. **(C)** The autocrine ligand-receptor interactions in neutrophil and macrophage. **(D)** The expression levels of ligands (CCL3/4) and the receptor (CCR5) in NF, DCM, and ICM.

ECM organization. ECM plays a vital role in cardiac homeostasis, which provides structural support for cardiac cells and maintains integrity and function by transducing important signals among different cells (Frangogiannis, 2019). The transformation of ECM patterns in biochemical in failing hearts hinged on the type of underlying injury

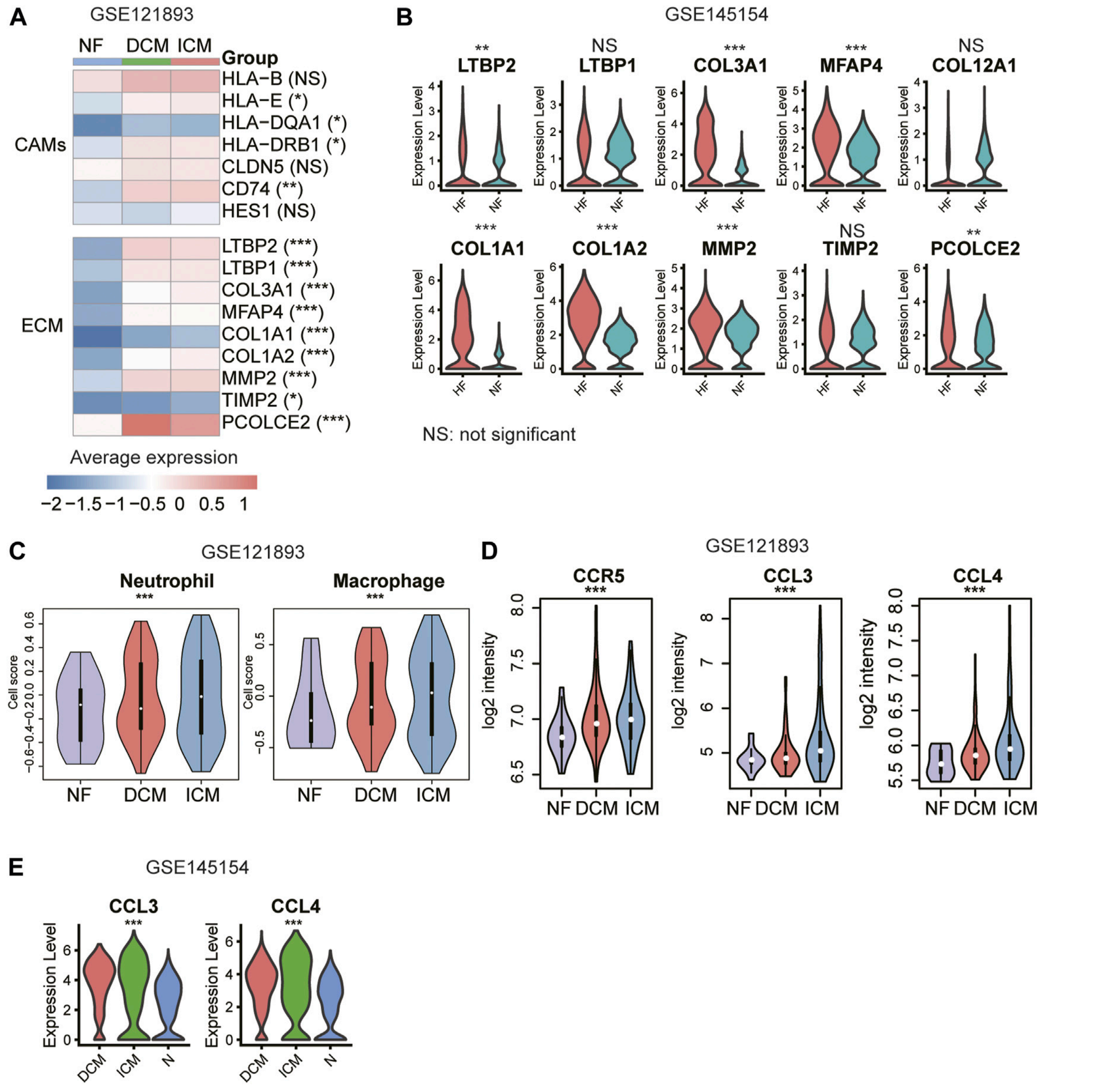


FIGURE 6 | Validation of the cell adhesion molecules (CAMs), extracellular matrix (ECM) genes, and immune responses. **(A)** The upregulation of CAMs and ECM genes in HF samples. **(B)** The differential expression levels of ECM genes between the fibroblasts of NF and HF (scRNA-seq dataset: GSE145154). **(C)** The higher abundance of neutrophil and macrophage in ICM. **(D)** The higher expression levels of CCR5, CCL3, and CCL4 in ICM. **(E)** The differential expression levels of CCL3 and CCL4 between the macrophages of NF, DCM and ICM (scRNA-seq dataset: GSE145154).

(Travers, Kamal, Robbins, Yutzey, & Blaxall, 2016). Collectively, our analysis confirmed that inflammation-related CAMs and ECM proteins such as collagens were specifically secreted by EC and fibroblast, respectively, and might induce cardiac inflammation and fibrosis during the progression of HF.

Previous studies have suggested that inflammation is a key factor of cardiovascular disease, with immune cell types such as macrophages and T lymphocytes mediating essential crosstalk in the progression to HF(Abplanalp et al., 2020). Since we found ICM had more specific immune cell types, such as macrophage and DC, we then focused on the activities of immune cells including macrophage and neutrophil. The cell-cell communication analysis revealed that the autocrine ligand-receptor interaction induced chemokine signaling activation in neutrophil and macrophage might be responsible for the immune response in ICM. During the process of cardiac inflammation, immune cells invade the cardiac tissue and coordinate the responses of damaging. Due to the length limitation of this article, we cannot describe all genes in detail. Taken together, our results suggested that higher inflammation in ICM might be associated with autocrine CCL3/CCL4-CCR5 interaction induced chemokine signaling activation. Furthermore, neutrophil and macrophage also appeared to be higher in ICM compared with DCM.

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

BH, TH, and L-hS conceived and designed the project and are responsible for the overall content. LZ, YL, H-wN, FL, and JX analyzed and interpreted the data. XS, LZ, and BH prepared the manuscript. XW, ZC, and L-hS contributed to revising the manuscript. All authors contributed to and discussed the results and critically reviewed the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to acknowledge funding from the National Natural Science Foundation of China (81900280, 81830010, 81330006, and 31701151), Shanghai Sailing Program (19YF1431600), Science and Technology Commission of Shanghai Municipality (18411950400), National Key R&D Program of China (2018YFC0910403), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), and Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200, XDA26040304).

REFERENCES

Abplanalp, W. T., John, D., Cremer, S., Assmus, B., Dorsheimer, L., Hoffmann, J., et al. (2020). Single-cell RNA-Sequencing Reveals Profound Changes in Circulating Immune Cells in Patients with Heart Failure. *Cardiovasc. Res.* 117, 484–494. doi:10.1093/cvr/cvaa101

Akhurst, R. J., and Hata, A. (2012). Targeting the TGFβ Signalling Pathway in Disease. *Nat. Rev. Drug Discov.* 11 (10), 790–811. doi:10.1038/nrd3810

Bacmeister, L., Schwarzl, M., Warnke, S., Stoffers, B., Blankenberg, S., Westermann, D., et al. (2019). Inflammation and Fibrosis in Murine Models of Heart Failure. *Basic Res. Cardiol.* 114 (3), 19. doi:10.1007/s00395-019-0722-5

Berezin, A. E., Kremzer, A. A., Martovitskaya, Y. V., Berezina, T. A., and Gromenko, E. A. (2016). Pattern of Endothelial Progenitor Cells and Apoptotic Endothelial Cell-Derived Microparticles in Chronic Heart Failure Patients with Preserved and Reduced Left Ventricular Ejection Fraction. *EBioMedicine* 4, 86–94. doi:10.1016/j.ebiom.2016.01.018

Bui, A. L., Horwich, T. B., and Fonarow, G. C. (2011). Epidemiology and Risk Profile of Heart Failure. *Nat. Rev. Cardiol.* 8 (1), 30–41. doi:10.1038/nrcardio.2010.165

Colombo, P. C., Banchs, J. E., Celaj, S., Talreja, A., Lachmann, J., Malla, S., et al. (2005). Endothelial Cell Activation in Patients with Decompensated Heart Failure. *Circulation* 111 (1), 58–62. doi:10.1161/01.CIR.0000151611.89232.3B

Davis, J., and Molkentin, J. D. (2014). Myofibroblasts: Trust Your Heart and Let Fate Decide. *J. Mol. Cell Cardiol.* 70, 9–18. doi:10.1016/j.yjmcc.2013.10.019

Frangogiannis, N. G. (2019). The Extracellular Matrix in Ischemic and Nonischemic Heart Failure. *Circ. Res.* 125 (1), 117–146. doi:10.1161/CIRCRESAHA.119.311148

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

Hannenhalli, S., Putt, M. E., Gilmore, J. M., Wang, J., Parmacek, M. S., Epstein, J. A., et al. (2006). Transcriptional Genomics Associates FOX Transcription Factors with Human Heart Failure. *Circulation* 114 (12), 1269–1276. doi:10.1161/CIRCULATIONAHA.106.632430

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* 14, 7. doi:10.1186/1471-2105-14-7

Humeres, C., and Frangogiannis, N. G. (2019). Fibroblasts in the Infarcted, Remodeling, and Failing Heart. *JACC: Basic Translational Sci.* 4 (3), 449–467. doi:10.1016/j.jacbts.2019.02.006

Jones, N. R., Roalfe, A. K., Adoki, I., Hobbs, F. D. R., and Taylor, C. J. (2019). Survival of Patients with Chronic Heart Failure in the Community: a Systematic Review and Meta-analysis. *Eur. J. Heart Fail.* 21 (11), 1306–1325. doi:10.1002/ejhf.1594

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: AnRPackage for Multivariate Analysis. *J. Stat. Soft.* 25 (1), 1–18. doi:10.18637/jss.v025.i01

Liu, Y., Morley, M., Brandimarto, J., Hannenhalli, S., Hu, Y., Ashley, E. A., et al. (2015). RNA-seq Identifies Novel Myocardial Gene Expression Signatures of Heart Failure. *Genomics* 105 (2), 83–89. doi:10.1016/j.ygeno.2014.12.002

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Ma, Y., Iyer, R. P., Jung, M., Czubryt, M. P., and Lindsey, M. L. (2017). Cardiac Fibroblast Activation Post-Myocardial Infarction: Current Knowledge Gaps. *Trends Pharmacol. Sci.* 38 (5), 448–458. doi:10.1016/j.tips.2017.03.001

Martini, E., Kunderfranco, P., Peano, C., Carullo, P., Cremonesi, M., Schorn, T., et al. (2019). Single-Cell Sequencing of Mouse Heart Immune Infiltrate in Pressure Overload-Driven Heart Failure Reveals Extent of Immune Activation. *Circulation* 140 (25), 2089–2107. doi:10.1161/CIRCULATIONAHA.119.041694

Moore-Morris, T., Guimarães-Camboa, N., Yutzey, K. E., Pucéat, M., and Evans, S. M. (2015). Cardiac Fibroblasts: from Development to Heart Failure. *J. Mol. Med.* 93 (8), 823–830. doi:10.1007/s00109-015-1314-y

Oatmen, K. E., Cull, E., and Spinale, F. G. (2020). Heart Failure as Interstitial Cancer: Emergence of a Malignant Fibroblast Phenotype. *Nat. Rev. Cardiol.* 17 (8), 523–531. doi:10.1038/s41569-019-0286-y

Oneglia, A., Nelson, M. D., and Merz, C. N. B. (2020). Sex Differences in Cardiovascular Aging and Heart Failure. *Curr. Heart Fail. Rep.* 17, 409–423. doi:10.1007/s11897-020-00487-7

Rao, M., Wang, X., Guo, G., Wang, L., Chen, S., Yin, P., et al. (2021). Resolving the Intertwining of Inflammation and Fibrosis in Human Heart Failure at Single-Cell Level. *Basic Res. Cardiol.* 116 (1), 55. doi:10.1007/s00395-021-00897-1

Schafer, S., Viswanathan, S., Widjaja, A. A., Lim, W.-W., Moreno-Moral, A., DeLaughter, D. M., et al. (2017). IL-11 Is a Crucial Determinant of Cardiovascular Fibrosis. *Nature* 552 (7683), 110–115. doi:10.1038/nature24676

Shantsila, E., Wrigley, B. J., Blann, A. D., Gill, P. S., and Lip, G. Y. H. (2012). A Contemporary View on Endothelial Function in Heart Failure. *Eur. J. Heart Fail.* 14 (8), 873–881. doi:10.1093/eurjhf/hfs066

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. 2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7), 1888–1902. doi:10.1016/j.cell.2019.05.031

Sweet, M. E., Cocciolo, A., Slavov, D., Jones, K. L., Sweet, J. R., Graw, S. L., et al. (2018). Transcriptome Analysis of Human Heart Failure Reveals Dysregulated Cell Adhesion in Dilated Cardiomyopathy and Activated Immune Pathways in Ischemic Heart Failure. *BMC Genomics* 19 (1), 812. doi:10.1186/s12864-018-5213-9

Travers, J. G., Kamal, F. A., Robbins, J., Yutzey, K. E., and Blaxall, B. C. (2016). Cardiac Fibrosis. *Circ. Res.* 118 (6), 1021–1040. doi:10.1161/CIRCRESAHA.115.306565

Triposkiadis, F., Xanthopoulos, A., and Butler, J. (2019). Cardiovascular Aging and Heart Failure. *J. Am. Coll. Cardiol.* 74 (6), 804–813. doi:10.1016/j.jacc.2019.06.053

Triposkiadis, F., Xanthopoulos, A., Parissis, J., Butler, J., and Farmakis, D. (2020). Pathogenesis of Chronic Heart Failure: Cardiovascular Aging, Risk Factors, Comorbidities, and Disease Modifiers. *Heart Fail. Rev.* 1, 1. doi:10.1007/s10741-020-09987-z

Vigil-Garcia, M., Demkes, C. J., Eding, J. E. C., Versteeg, D., de Ruiter, H., Perini, I., et al. (2020). Gene Expression Profiling of Hypertrophic Cardiomyocytes Identifies New Players in Pathological Remodelling. *Cardiovasc. Res.* 117, 1532–1545. doi:10.1093/cvr/cvaa233

Wang, L., Yu, P., Zhou, B., Song, J., Li, Z., Zhang, M., et al. (2020). Single-cell Reconstruction of the Adult Human Heart during Heart Failure and Recovery Reveals the Cellular Landscape Underlying Cardiac Function. *Nat. Cel Biol* 22 (1), 108–119. doi:10.1038/s41556-019-0446-7

Yamaguchi, T., Sumida, T. S., Nomura, S., Satoh, M., Higo, T., Ito, M., et al. (2020). Cardiac Dopamine D1 Receptor Triggers Ventricular Arrhythmia in Chronic Heart Failure. *Nat. Commun.* 11 (1), 4364. doi:10.1038/s41467-020-18128-x

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Ziaeian, B., and Fonarow, G. C. (2016). Epidemiology and Aetiology of Heart Failure. *Nat. Rev. Cardiol.* 13 (6), 368–378. doi:10.1038/nrcardio.2016.25

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 22 February 2022
doi: 10.3389/fbioe.2022.829509

Spatially Annotated Single Cell Sequencing for Unraveling Intratumor Heterogeneity

Myrthe M. Smit[1,2], *Kate J. Feller*[1,2], *Li You*[1,2], *Jelle Storteboom*[1,2], *Yasin Begce*[1,2], *Cecile Beerens*[1,2] *and Miao-Ping Chien*[1,2,3]*†

[1]*Department of Molecular Genetics, Erasmus University Medical Center, Rotterdam, Netherlands,* [2]*Erasmus MC Cancer Institute, Rotterdam, Netherlands,* [3]*Oncode Institute, Utrecht, Netherlands*

OPEN ACCESS

Edited by:
Zhaoyuan Fang,
Zhejiang University, China

Reviewed by:
Jianhua Xing,
University of Pittsburgh, United States
Shrabanti Chowdhury,
Icahn School of Medicine at Mount Sinai, United States
Salim Ghannoum,
University of Oslo, Norway

****Correspondence:***
Miao-Ping Chien
m.p.chien@erasmusmc.nl

†*These authors share senior authorship*

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *05 December 2021*
Accepted: *11 January 2022*
Published: *22 February 2022*

Citation:
Smit MM, Feller KJ, You L, Storteboom J, Begce Y, Beerens C and Chien M-P (2022) Spatially Annotated Single Cell Sequencing for Unraveling Intratumor Heterogeneity. Front. Bioeng. Biotechnol. 10:829509. doi: 10.3389/fbioe.2022.829509

Intratumor heterogeneity is a major obstacle to effective cancer treatment. Current methods to study intratumor heterogeneity using single-cell RNA sequencing (scRNA-seq) lack information on the spatial organization of cells. While state-of-the art spatial transcriptomics methods capture the spatial distribution, they either lack single cell resolution or have relatively low transcript counts. Here, we introduce spatially annotated single cell sequencing, based on the previously developed functional single cell sequencing (FUNseq) technique, to spatially profile tumor cells with deep scRNA-seq and single cell resolution. Using our approach, we profiled cells located at different distances from the center of a 2D epithelial cell mass. By profiling the cell patch in concentric bands of varying width, we showed that cells at the outermost edge of the patch responded strongest to their local microenvironment, behaved most invasively, and activated the process of epithelial-to-mesenchymal transition (EMT) to migrate to low-confluence areas. We inferred cell-cell communication networks and demonstrated that cells in the outermost ~10 cell wide band, which we termed the invasive edge, induced similar phenotypic plasticity in neighboring regions. Applying FUNseq to spatially annotate and profile tumor cells enables deep characterization of tumor subpopulations, thereby unraveling the mechanistic basis for intratumor heterogeneity.

Keywords: spatial transcriptomics, single cell sequencing, functional single cell sequencing, intratumoral heterogeneity, epithelial-to-mesenchym transition (EMT)

INTRODUCTION

Intratumor heterogeneity, both at the genetic and transcriptomic level, is commonly observed in various cancer types and complicates diagnosis and treatment (Gerlinger et al., 2012; Patel et al., 2014; Morrissy et al., 2017; Puram et al., 2017; Berglund et al., 2018). Rare populations of cells can contribute to increased tumor progression (Burrell et al., 2013; Patel et al., 2014), metastatic potential (Yachida et al., 2010; Navin et al., 2011) and therapy resistance (Sottoriva et al., 2013; Patel et al., 2014; Tirosh et al., 2016). Single-cell sequencing is key to characterizing the complexity of intratumor heterogeneity, but lacks information about functional properties and spatial organization of cells (Lawson et al., 2018). We have recently developed a functionally annotated transcriptomic profiling technique, called functional single cell sequencing (FUNseq), to study heterogeneous populations of tumor cells based on functional features (You et al., 2021). This technology uses live-cell imaging to identify cells with a phenotype of interest (e.g., cell migration or morphology), which can then be

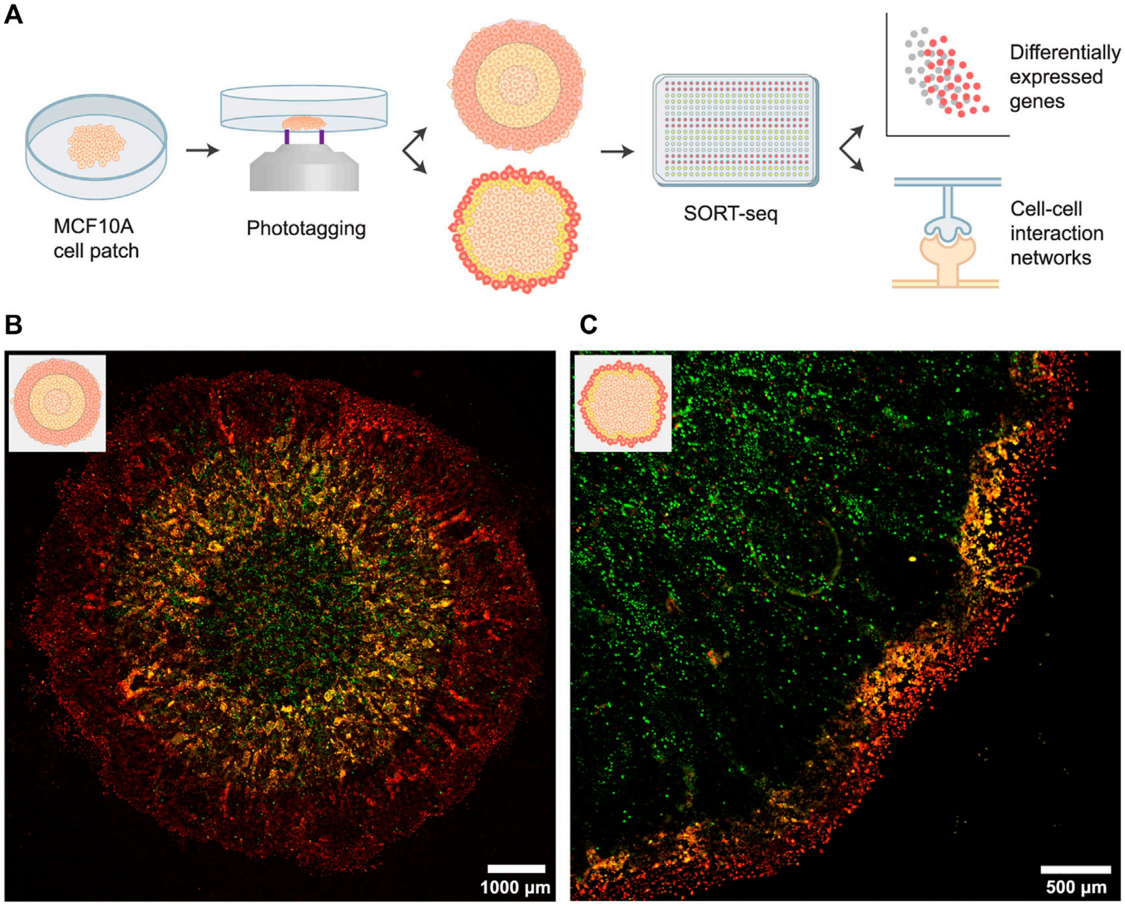

A
MCF10A cell patch
Phototagging
SORT-seq
Differentially expressed genes
Cell-cell interaction networks
B
1000 µm
C
500 µm

FIGURE 1 | Spatially profiling an *in vitro* tumor model using the FUNseq technology. **(A)** Schematic depiction of the assay, cell labeling and scRNA-seq analysis. For the cell labeling (middle panel), we either phototagged concentric rings of equal width (top; 1,000–1,500 µm bandwidth) or 250 µm wide bands at the invasive edge (bottom). In both approaches, the outer population was labeled with JF646 phototagging dye (red) and the middle population was labeled with both JF549 and JF646 (yellow). **(B)** Patch of MCF10A cells expressing a GFP marker that was phototagged with the larger bandwidth. Green: GFP, yellow: JF549, red: JF646. **(C)** Phototagging the invasive edge of a MCF10A cell patch yields well-demarcated bands of cells.

phototagged (*via* a photopatterned device) with a photoactivatable dye, isolated and subjected to single-cell RNA sequencing (scRNA-seq). Hence, FUNseq links phenotypic traits to gene expression profiles of rare subpopulations of tumor cells, thereby identifying the underlying mechanisms of intratumor heterogeneity. However, cells are currently labeled using a single dye, making it impossible to discern cells based on their spatial organization.

Here, we applied FUNseq to characterize intratumor heterogeneity in tumor subpopulations that are spatially located differently in an untransformed, mammary epithelial tumor model. We specifically focused on the epithelial-to-mesenchymal transition (EMT), as this is an important source for intratumor heterogeneity (Nieto et al., 2016). During EMT, epithelial cells gradually acquire a mesenchymal phenotype, thereby losing their cell-cell adhesion and cell polarity while gaining the ability to migrate and invade (Nieto et al., 2016; Pastushenko et al., 2018; Revenco et al., 2019). EMT can be induced by multiple stimuli, including various growth factors and a cell's local microenvironment (Cook and Vanderhyden 2020). To illustrate, cells at the migrating front of tumors express higher levels of EMT marker genes than cells in the center (Puram et al., 2017). Recently, McFaline-Figueroa et al. (2019) made a similar observation using an *in vitro* tumor model, showing that untransformed MCF10A cells in the outer layer of a high-confluence patch of cells undergo EMT. However, the exact transcriptomic changes that cause this EMT are currently unknown. To identify the genes that drive the outward migration, one needs to profile the cells in the outermost layer of the cell patch (i.e., the invasive edge). This could be done by spatially annotating bands of cells before subjecting them to scRNA-seq, which enables specific characterization of the invasive edge.

Using a similar tumor model as McFaline-Figueroa et al., we applied FUNseq to profile MCF10A epithelial cells that were spatially located in the outer layer (~1,000–1,500 µm bandwidth, ~50 cell wide band) or the outermost layer (250 µm bandwidth, ~10 cell wide band) of the cell mass. We demonstrated that cells in the outermost layer were progressing through EMT and induced similar phenotypic plasticity in neighboring regions.

Using cell-cell communication network analysis, we also showed that the Ephrin, EGF and VEGF signaling pathways were involved in driving this invasive behavior. Our data indicates that FUNseq can spatially profile intratumor heterogeneity, thereby unraveling the underlying mechanisms for the observed phenotypic variations.

RESULTS

FUNseq Can Spatially Annotate and Profile Cells With Desired Spatial Bandwidths

We applied FUNseq to profile spatial heterogeneity in an *in vitro* tumor model: untransformed, mammary epithelial MCF10A cells (**Figure 1A**). MCF10A cells expressing a GFP marker were seeded in a high-confluence, circular patch. After growing the cells for 6 days, cells at the leading edge of the patch acquired a spindle-like morphology and migrated to unoccupied areas of the dish (**Supplementary Figure S1**), indicating that they might have undergone EMT (Vuoriluoto et al., 2011).

Next, we imaged the cells using our custom-built Ultrawide Field-of-view Optical (UFO) microscope (You et al., 2021) and identified the outer, middle and inner regions (with a bandwidth of 1,000–1,500 μm) of the patch. Cells were first incubated with photoactivatable Janelia Fluor 646 (JF646) dye, after which we phototagged the outer one-third of the patch (**Figure 1B**; cells emit red fluorescence (λ_{ex}: ~650 nm, λ_{em}: ~665 nm) after photoactivation). Subsequently, we incubated cells with photoactivatable Janelia Fluor 549 dye (JF549) and phototagged the middle one-third of the patch (cells emit green fluorescence (λ_{ex}: ~550 nm, λ_{em}: ~570 nm) after photoactivation). Hence, cells in the middle ring were labeled with both dyes, as the cytoplasmic JF646 dye is retained within cells. Labeled populations were isolated by flow cytometry and sequenced using SORTseq, a plate-based, modified CEL-seq2 scRNA-seq technology (Hashimshony et al., 2016; Muraro et al., 2016).

A similar labeling strategy can be used to profile the invasive edge at a higher resolution. For this, we phototagged cells in the outermost layer (250 μm bandwidth, ~10 cell wide band) of the patch with JF646 and we phototagged cells in the next 250 μm with both JF549 and JF646. Live-cell imaging of the labeled patches showed well-demarcated bands of cells (**Figure 1C**), validating that FUNseq can be used to annotate and isolate confined tumor regions with desired spatial bandwidth.

FUNseq Identified Subtle Variations in Gene Expression Profiles Between Tumor Regions

To couple the observed phenotypic plasticity in the outer layer to underlying transcriptomic changes, tumor subpopulations were subjected to scRNA-seq. We sequenced two biological replicates of patches phototagged with the larger bandwidth, yielding a total of 743 analyzed single cell transcriptomes (**Supplementary Figure S2**). Dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) indicated a modest separation of the populations but did not form coherent clusters (**Figure 2A**), suggesting that there is substantial similarity of the gene expression profiles between the tumor regions.

To quantity the level of EMT in each subpopulation, we calculated EMT scores using Gene Set Variation Analysis (GSVA) (Hänzelmann et al., 2013). For each cell, an epithelial (E) and mesenchymal (M) score was calculated using two gene sets containing 65 epithelial and 115 mesenchymal genes (Cesano 2015). Following the approach of Sacchetti et al. (2021), we subtracted the E score from the M score to define a single EMT score for each cell (EMT = M − E). Cells in the outer layer had significantly higher EMT scores than cells in the center (Kruskal-Wallis test, p = .0017; **Figure 2B**). However, no significant changes between adjacent populations were observed, presumably because the relatively large number of cells per region led to substantial heterogeneity within each population (**Supplementary Figure S3**). This solidified our notion that one needs to specifically profile the invasive edge to reliably identify the transcriptomic drivers for migration and invasion. Hence, we next sought to profile the migrating front at a higher resolution.

Cells at the Invasive Edge Strongly Activated the Epithelial-to-Mesenchymal Transition

We phototagged the migrating front (~10 cell wide bands) and separated the outermost cells from the inner tumor mass (**Figure 1C**). Using this high-resolution phototagging approach, we analyzed 696 single cell transcriptomes from two biological replicates. Dimensionality reduction now revealed coherent clusters of cells that segregate based on the spatial populations (**Figure 2C**). The middle and outermost layer clustered together in the UMAP embedding, presumably since cells in both layers are progressing through EMT. Classically, EMT has been viewed as a discrete process in which cells pass through distinct transition stages before acquiring a fully mesenchymal morphology (Pastushenko and Blanpain 2019). Our UMAP embedding (**Figure 2D**) indicated that EMT scores vary continuously across the embedded cells, further solidifying recent findings that EMT is a continuous process (McFaline-Figueroa et al., 2019; Cook and Vanderhyden 2020). Expression of the classic epithelial markers E-cadherin (CDH1) and EPCAM gradually decreased from the center to the edge of the patch, while the mesenchymal markers VIM and FN1 showed a reciprocal pattern, suggesting that cells are exhibiting epithelial-mesenchymal plasticity (Zhao et al., 2015; Yang et al., 2020) (**Figure 2E**; **Supplementary Figure S4**). These changes in CDH1 expression were not detected by McFaline-Figueroa et al. (2019), underscoring the value of deep sequencing using FUNseq to resolve subtle transcriptomic changes. Moreover, we found that adjacent populations have significantly varying EMT scores ($p <$ .0001; **Figure 2F**), further increasing our confidence that profiling the invasive edge of the tumor model could identify drivers of migration and invasion.

Next, we identified differentially expressed genes (DEGs) between the subpopulations and found that classic

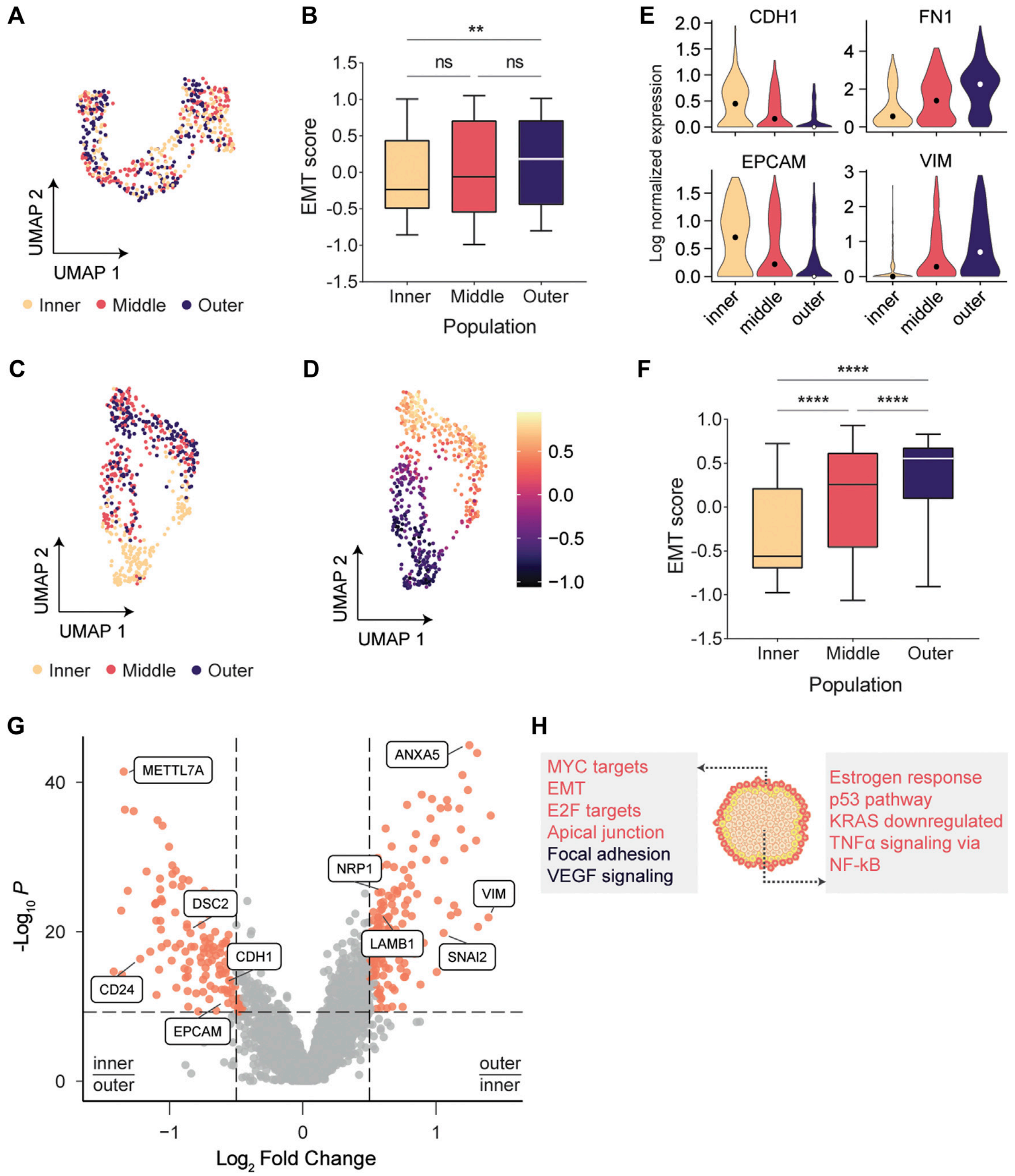


FIGURE 2 | scRNA-seq indicated that cells at the invasive edge were progressing through EMT. **(A)** UMAP embedding of cells labeled with the larger bandwidth showed a modest separation of tumor regions, but no coherent clusters were formed. **(B)** EMT scores between inner and outer populations vary significantly ($p = .0017$; Kruskal-Wallis test). **(C)** Inner and outermost tumor regions labeled with the smaller bandwidth separate clearly in UMAP space. **(D)** EMT scores gradually increase across the UMAP embedding. **(E)** Expression of classic epithelial markers decreases radially outwards while expression of classic mesenchymal markers increases. **(F)** EMT scores are significantly varying between adjacent populations ($p < .0001$; Kruskal-Wallis test). **(G)** Volcano plot indicating genes overexpressed in the outermost population ($\log_2(FC) > .5$) and in the inner population ($\log_2(FC) < -.5$). **(H)** Overrepresentation analysis using the MSigDB Hallmarks (red) and Wikipathways (black) databases.

mesenchymal markers such as VIM and the EMT transcription factor SNAI2 were upregulated in the outermost population, while the epithelial markers CDH1 and EPCAM were upregulated in the center of the patch (**Figure 2G**; **Supplementary Figure S4**). Genes upregulated in the outermost ($n = 165$ DEGs) or center ($n = 142$ DEGs) population were used for overrepresentation analysis using the MSigDB Hallmarks gene set collection (Liberzon et al., 2015) and the Wikipathways database (Martens et al., 2021) (**Figure 2H**). As expected, genes involved in EMT and extracellular matrix

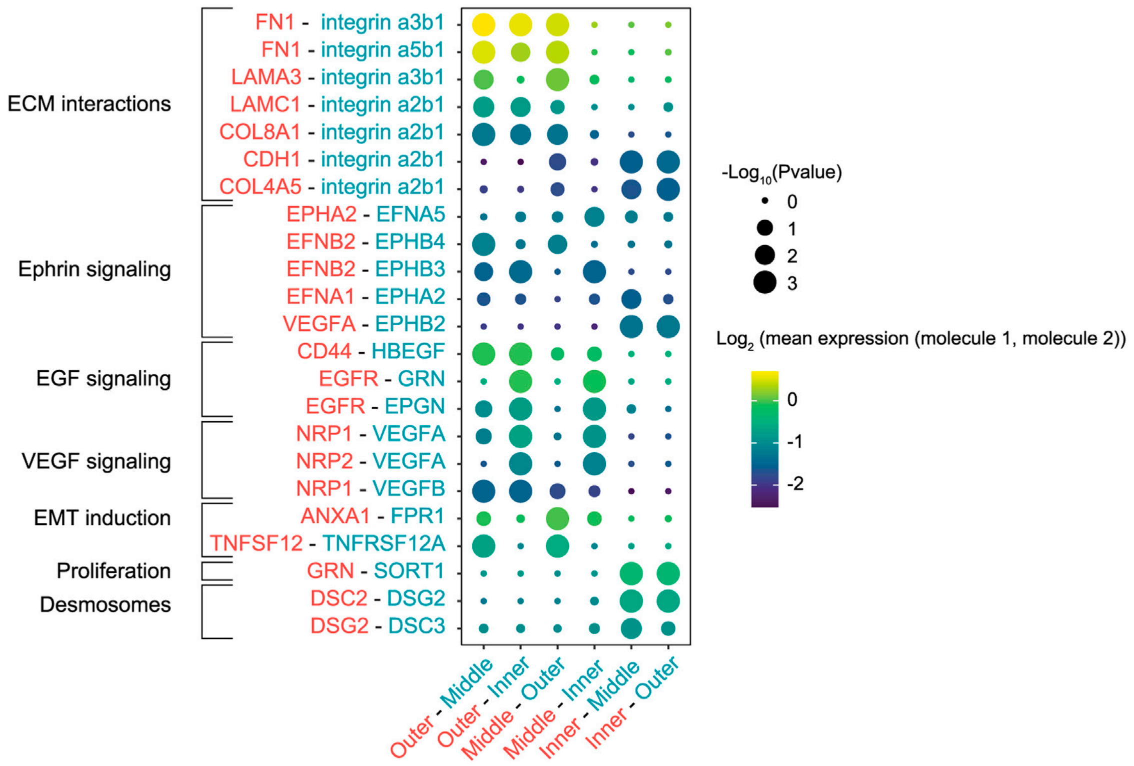


FIGURE 3 | Cell-cell interactions between cells in various patch regions labeled with the smaller bandwidth. Interactions were inferred based on the expression of ligands and receptors in the different cell populations. The first molecule in each interaction pair (rows) corresponds to the first region in each population pair (columns). Circles scaled by the significance of the interaction and colored by the average expression level of ligand and receptor.

interactions were overrepresented in the outermost population. Additionally, cells at the invasive edge were enriched for the vascular endothelial growth factor (VEGF) signaling pathway, which can induce cell migration and EMT (Anthony D. Yang et al., 2006; Gonzalez-Moreno et al., 2010; Bhattacharya et al., 2017). VEGF can activate the neuropilin-1 receptor (NRP1), which is upregulated in the outermost population (**Figure 2G**; **Supplementary Figure S5**) and promotes proliferation, migration and invasion of tumor cells (Goel and Mercurio 2013; Luo et al., 2016).

Cell-Cell Communication Network Analysis Identified Multiple Epithelial-to-Mesenchymal Transition Inducers

Since a wide range of transcription factors and extracellular stimuli are involved in stimulating EMT (Nieto et al., 2016), we next set out to map the cell-cell communication networks that regulate the EMT in our *in vitro* tumor model. We re-analyzed scRNA-seq data from the high-resolution labeling experiment to identify interactions between the different populations using CellPhoneDB, a repository of ligand-receptor complexes that can predict enriched cellular interactions based on the expression of ligands and receptors in cell populations (Efremova et al., 2020). The outermost population was highly enriched for fibronectin (FN1), laminin (LAMA3 and LAMC1) and collagen (COL8A1 and COL4A1) expression, extracellular matrix (ECM) proteins that can interact with the integrins expressed in the middle and inner populations (**Figure 3**). Specifically, interactions of fibronectin and laminin with the α3β1 integrin modulate cell adhesion to the ECM and cell motility (Meng et al., 2009; Hamill et al., 2010; Jia et al., 2010; Zhang et al., 2017). Interestingly, this analysis predicted multiple interactions in the Ephrin-signaling pathway, in which ligands and receptors activate bidirectional signals that can lead to somewhat paradoxical downstream effects (Pasquale 2008). To illustrate, cells in the outermost and middle populations expressed the EphB4 receptor and its ligand EphrinB2 (EFNB2) (**Supplementary Figure S6**). Activation of EphB4 induces cell migration and invasion in cancer cells (Steinle et al., 2002; Kumar et al., 2006; Nai-Ying Yang et al., 2006), although the exact opposite effect has also been reported (Noren et al., 2006). Additionally, reverse signaling through EphrinB2 can stimulate cell migration through the PI3K pathway (Steinle et al., 2002; Kumar et al., 2006).

Finally, CellPhoneDB inferred enrichment of multiple EMT inducers and their receptors in the outermost population, such as tumor necrosis factor (TNFA) and genes involved in the EGF pathway (CD44, EGFR, EPGN, HBEGF) (Cheng et al., 2012; Revenco et al., 2019; Cook and Vanderhyden 2020). Conversely, cells in the center of the patch were enriched for DSC2 and DSG2, genes that encode components of desmosome cell-cell junctions (Garrod and Chidgey 2008; Nekrasova and Green 2013), hallmarks of an epithelial phenotype. Taken together, the identified cell-cell interactions indicated that cells at the migrating front responded to their local microenvironment and stimulated similar invasive behavior in neighboring regions.

DISCUSSION

Intratumor heterogeneity is a major challenge for effective cancer treatment. Single-cell genomics and transcriptomics proved themselves valuable methods to study this heterogeneity, but lack information about the spatial organization of cells. Recently, various spatial transcriptomics methods have been developed to add positional information from tissue sections to the single-cell transcriptomes (Ståhl et al., 2016; Vickovic et al., 2019; Stickels et al., 2021), providing numerous insights in cancer biology and other fields (Longo et al., 2021). However, these methods either lack single cell resolution or have substantially lower transcript counts per cell than conventional scRNA-seq. Here, we applied our recently developed FUNseq technology to spatially profile confined tumor regions. The strength of this method lies in the combination of labeling tumor regions guided by live-cell imaging and deep sequencing of single cells. This allowed us to profile gene expression in isolated tumor regions using 34,000 transcripts per cell (**Supplementary Figure S2**), compared to the 494 and 11.5 transcripts per 10 μm bead for Slide-seq V2 and HDST, respectively (Stickels et al., 2021). The increased sensitivity of FUNseq allows us to study low abundance transcripts, enabling deep characterization of tumor cells.

We profiled tumor heterogeneity in an *in vitro* tumor model (McFaline-Figueroa et al., 2019) by annotating cells located at different distances from the center of a 2D epithelial cell mass. Cells in the outermost layer or invasive edge (~10 cell wide band) of this patch were progressing through EMT, suggesting that these cells sense their local microenvironment and acquire a mesenchymal phenotype to migrate to unoccupied areas of the dish. Taking advantage of the FUNseq's deep sequencing, we characterized cell-cell interaction networks between the different tumor regions. We identified various interactions between outermost cells and ECM components that can stimulate cell migration and we showed that outermost cells are enriched for ligands and receptors that can stimulate EMT, such as components of the Ephrin, EGF and VEGF signaling pathways.

By combining phototagging of confined tumor regions and deep sequencing of single cells, we characterized the transcriptomic heterogeneity in a population of untransformed epithelial cells. To fully explore the potential of FUNseq, the next step would be to profile tumor sections, which have much higher complexity than relatively homogeneous cell lines. We envision that FUNseq might address important questions about intratumor heterogeneity, such as how tumor cells interact with the tumor microenvironment and how tumor composition affects treatment outcome.

In summary, we demonstrated that FUNseq can spatially annotate and profile subpopulations of an *in vitro* tumor model. We showed that cells at the invasive edge (~10 cells wide band) of a high-confluence patch of cells underwent EMT, migrated to low-confluence areas and induced similar phenotypic plasticity in neighboring cells. Spatially profiling tumor cells using FUNseq enables deep characterization of intratumor heterogeneity, thereby laying the foundation for a more complete understanding of tumor biology.

MATERIALS AND METHODS

Cell Culture

MCF10A_H2B_GFP human breast epithelial cells were a kind gift of Reuven Agami (Netherlands Cancer Institute). Cells were cultured at 37°C and 5% CO_2 in DMEM/F12 medium without phenol red (Gibco), supplemented with 5% Donor Equine Serum, 1% penicillin/streptomycin, 20 ng/ml EGF, 500 ng/ml hydrocortisone, 100 ng/ml cholera toxin and 10 μg/ml insulin.

Before conducting experiments, cells were seeded on 20 mm glass bottom dishes (Cellvis), coated with 0.1 mg/ml fibronectin (EMD Millipore). 10,000 cells were seeded in a droplet in the center of the dish, such that a circular patch of cells was formed in the center of the dish. After 4.5 h, dishes were washed with Dulbecco's PBS (Sigma) to remove non-adherent cells. The patch of cells was then cultured in MCF10A medium at 37°C and 5% CO_2 for 6 days.

Imaging and Cell Labeling

Cell labeling was performed on the Ultrawide Field-of-view Optical (UFO) microscope developed previously (You et al., 2021). Cells were incubated with 40 μM photoactivatable Janelia Fluor 646 (JF646) dye (Tocris) for 20 min and washed with MCF10A culture medium. Bright-field images were used to localize the patch of cells, after which we identified the cells to be labeled using a low-resolution or high-resolution approach. In the low-resolution tagging approach, we fit three concentric rings with equal bandwidth (1,000–1,500 μm bandwidth) in the area of the patch. In the high-resolution approach, we divide the patch of cells in three layers: the outermost 250 μm of cells (~10 cell wide band), the next 250 μm, and the inside of the patch.

In both approaches, the outer population of cells was then selectively illuminated for 2 min with 405 nm light using a digital micromirror device (DMD), thereby phototagging these cells with JF646. Next, cells were incubated with 40 μM photoactivatable Janelia Fluor 549 (JF549) dye (Grimm et al., 2016) (Tocris) for 20 min and washed with MCF10A culture medium. The imaging and labeling process was repeated, but now illuminating the middle population of cells. These cells are thus phototagged with JF549 and JF646, as both dyes are present in the cytoplasm and become activated upon illumination. For visualization purposes, image background was subtracted and image contrast was adjusted using ImageJ.

Cell Isolation and Single-Cell RNA Sequencing

Cells were harvested using trypsin-EDTA without phenol red (Sigma), centrifuged and resuspended in HBSS buffer (Gibco). Live single cells (validated by Draq7 viability staining) were sorted into 384-well plates using the BD FACSMelody Cell Sorter (BD Biosciences), spun-down and stored at −80°C. Library preparation and single-cell RNA sequencing was performed by Single Cell Discoveries (Utrecht, Netherlands) using their custom SORT-seq protocol (Muraro et al., 2016). cDNA libraries were sequenced at 150 k reads/cell on the Illumina NextSeq 500 platform.

scRNA-Seq Analysis

scRNA-seq data was aligned and preprocessed by Single Cell Discoveries as described by Muraro et al. (2016). Gene expression matrices were processed using Seurat v4 (Hao et al., 2021). Cells containing 2,000–9,000 features and less than 40% mitochondrial genes were selected. Gene expression was either normalized using the SCTransform (Hafemeister and Satija 2019) function for dimensionality reduction, or log-normalized for all other downstream analysis. Cell cycle scoring and regression was performed using a set of G2/M and S phase markers (Tirosh et al., 2016). We performed a Principal Component Analysis (PCA) on the normalized gene expression data and used the first 40 principal components for dimensionality reduction using UMAP.

Differentially expressed genes between the inner and outer populations were identified with Seurat's findMarkers function using a Wilcoxon rank-sum test and filtering for genes with a Bonferroni corrected p-value $< 1 \times 10^{-5}$. Genes with $\log_2$ fold change >0.5 were marked as upregulated and genes with $\log_2$ fold change <−0.5 were marked as downregulated. Next, overrepresentation analysis (ORA) was performed with the ClusterProfiler v4 package (Wu et al., 2021). The enricher function was used with default settings (one-sided Fisher's exact test with Benjamini-Hochberg adjusted p-values) and the most significantly enriched processes were visualized.

To calculate the level of EMT in each cell, we followed the approach of Sacchetti et al. (2021) Gene Set Variation analysis was performed using the GSVA package (Hänzelmann et al., 2013), where we used a set of EMT markers that is publicly available from the Nanostring nCounter PanCancer Progression Panel (Cesano 2015). This gene set contained 65 epithelial (E) and 115 mesenchymal (M) genes (**Supplementary Table S1**). For each cell we calculated its GSVA enrichment scores for the epithelial and mesenchymal genes, after which we subtracted the E score from the M score to define the cell's EMT score.

Enriched ligand-receptor interactions between the different populations of cells were inferred using the CellphoneDB package (Efremova et al., 2020). This analysis uses empirical shuffling to identify enriched ligand-receptor interactions based on the expression levels in the different populations, while requiring that all subunits from heteromeric ligand-receptor complexes are expressed. Log-normalized gene expression matrices were used as input files and the statistical analysis (without subsampling) was performed using a p-value threshold of .01 and requiring that at least 20% of the cells in a population expresses a specific ligand-receptor interaction. To identify highly specific interactions between populations, we filtered for interactions with rank ≤.444. In this way, we filtered for ligand-receptor interactions that were significantly enriched in ≤4 population pairs (out of 9 population pairs in our setup). After this initial prioritization of the predicted interactions, we manually selected biologically relevant interactions for visualization.

DATA AVAILABILITY STATEMENT

TThe raw data used to generate for this study can be found at NCBI's GEO DataSets site with an ID number of GSE196245.

AUTHOR CONTRIBUTIONS

MS conducted the experiments and performed the single cell RNA sequencing and CellPhoneDB analysis. KF conducted part of the experiments and scripted the single cell RNA sequencing analysis code. LY scripted the photopatterned code and performed some of the image analysis. JS upgraded the microscope and program for the spatial phototagging experiment. YB helped with the whole FUNseq pipeline and cell sorting. CB contributed to part of the cell culture preparation for the experiments. MS, KF, and M-PC designed most of the experiments. MS and M-PC wrote the paper with input from all authors. M-PC contributed to and supervised all aspects of the project.

ACKNOWLEDGEMENTS

M-PC acknowledges support from the Oncode Institute, Cancer GenomiCs.nl (CGC), NWO (Netherlands Organization for Scientific Research) Veni Grant, Stichting Ammodo and Erasmus MC grant. M-PC appreciates Josephine Nefkens Stichting's support on the UFO microscope. We thank Reuven Agami for the kind gift of the MCF10A-H2B-GFP cell line.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.829509/full#supplementary-material

REFERENCES

Berglund, E., Jonas, M., Schultz, N., Friedrich, S., Marklund, M., Joseph, B., et al. (2018). Spatial Maps of Prostate Cancer Transcriptomes Reveal an Unexplored Landscape of Heterogeneity. *Nat. Commun.* 9, 2419. doi:10.1038/s41467-018-04724-5

Bhattacharya, R., Fan, F., Wang, R., Ye, X., Xia, L., Boulbes, D., et al. (2017). Intracrine VEGF Signalling Mediates Colorectal Cancer Cell Migration and Invasion. *Br. J. Cancer* 117 (6), 848–855. doi:10.1038/bjc.2017.238

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The Causes and Consequences of Genetic Heterogeneity in Cancer Evolution. *Nature* 501, 338–345. doi:10.1038/nature12625

Cesano, A. (2015). nCounter(®) PanCancer Immune Profiling Panel (NanoString Technologies, Inc., Seattle, WA). *J. Immunother. Cancer* 3, 42–43. doi:10.1186/s40425-015-0088-7

Cheng, J.-C., Auersperg, N., Leung, P. C. K., and Leung, K. (2012). EGF-induced EMT and Invasiveness in Serous Borderline Ovarian Tumor Cells: A Possible Step in the Transition to Low-Grade Serous Carcinoma Cells? *PLOS ONE* 7, e34071. doi:10.1371/journal.pone.0034071

Cook, D. P., and Vanderhyden, B. C. (2020). Context Specificity of the EMT Transcriptional Response. *Nat. Commun.* 11, 2142–2149. doi:10.1038/s41467-020-16066-2

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (20202020). CellPhoneDB: Inferring Cell-Cell Communication from Combined Expression of Multi-Subunit Ligand-Receptor Complexes. *Nat. Protoc.* 15 (4), 1484–1506. doi:10.1038/s41596-020-0292-x

Garrod, D., and Chidgey, M. (2008). Desmosome Structure, Composition and Function. *Biochim. Biophys. Acta (Bba) - Biomembranes* 1778, 572–587. doi:10.1016/j.bbamem.2007.07.014

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* 366, 883–892. doi:10.1056/nejmoa1113205

Goel, H. L., and Mercurio, A. M. (2013). VEGF Targets the Tumour Cell. *Nat. Rev. Cancer* 13 (12), 871–882. doi:10.1038/nrc3627

Gonzalez-Moreno, O., Lecanda, J., Green, J. E., Segura, V., Catena, R., Serrano, D., et al. (2010). VEGF Elicits Epithelial-Mesenchymal Transition (EMT) in Prostate Intraepithelial Neoplasia (PIN)-like Cells via an Autocrine Loop. *Exp. Cell Res* 316, 554–567. doi:10.1016/j.yexcr.2009.11.020

Grimm, J., English, B., and Choi, H. (2016). Bright Photoactivatable Fluorophores for Single-Molecule Imaging. *Nat. Methods* 13, 985–988. doi:10.1038/nmeth.4034

Hafemeister, C., and Satija, R. (2019). Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression. *Genome Biol.* 20 (1), 1–15. doi:10.1186/s13059-019-1874-1

Hamill, K. J., Paller, A. S., and Jones., J. C. (2010). Adhesion and Migration, the Diverse Functions of the Laminin Alpha3 Subunit. *Dermatol. Clin.* 28, 79–87. doi:10.1016/j.det.2009.10.009

Hänzelmann, S., Castelo, R., and Guinney, J. (20132013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* 14 (1), 7–15. doi:10.1186/1471-2105-14-7

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., et al. (2021). Integrated Analysis of Multimodal Single-Cell Data. *Cell* 184, 3573–3587. doi:10.1016/j.cell.2021.04.048

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., et al. (2016). CEL-Seq2: Sensitive Highly-Multiplexed Single-Cell RNA-Seq. *Genome Biol.* 17, 77–7. doi:10.1186/s13059-016-0938-8

Jia, D., Yan, M., Wang, X., Hao, X., Liang, L., Liu, L., et al. (2010). Development of a Highly Metastatic Model that Reveals a Crucial Role of Fibronectin in Lung Cancer Cell Migration and Invasion. *BMC Cancer* 10, 1–12. doi:10.1186/1471-2407-10-364

Kumar, S. R., Singh, J., Xia, G., Krasnoperov, V., Hassanieh, L., Ley, E. J., et al. (2006). Receptor Tyrosine Kinase EphB4 Is a Survival Factor in Breast Cancer. *Am. J. Pathol.* 169, 279–293. doi:10.2353/ajpath.2006.050889

Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., and Werb, Z. (2018). Tumour Heterogeneity and Metastasis at Single-Cell Resolution. *Nat. Cell Biol* 20 (12), 1349–1360. doi:10.1038/s41556-018-0236-7

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection. *Cell Syst* 1, 417–425. doi:10.1016/j.cels.2015.12.004

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

Luo, M., Hou, L., Li, J., Shao, S., Huang, S., Meng, D., et al. (2016). VEGF/NRP-1 axis Promotes Progression of Breast Cancer via Enhancement of Epithelial-Mesenchymal Transition and Activation of NF-Kb and β-catenin. *Cancer Lett.* 373, 1–11. doi:10.1016/j.canlet.2016.01.010

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., et al. (2021). WikiPathways: Connecting Communities. *Nucleic Acids Res.* 49, D613–D621. doi:10.1093/nar/gkaa1024

McFaline-Figueroa, J. L., Hill, A. J., Qiu, X., Jackson, D., Shendure, J., and Trapnell, C. (2019). A Pooled Single-Cell Genetic Screen Identifies Regulatory Checkpoints in the Continuum of the Epithelial-To-Mesenchymal Transition. *Nat. Genet.* 51, 1389–1398. doi:10.1038/s41588-019-0489-5

McInnes, L., Healy, J., and James, M. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *J. Open Source Softw.* 3, 861. doi:10.21105/joss.00861

Meng, X. N., Jin, Y., Yu, Y., Bai, J., Liu, G. Y., Zhu, J., et al. (2009). Characterisation of Fibronectin-Mediated FAK Signalling Pathways in Lung Cancer Cell Migration and Invasion. *Br. J. Cancer* 101 (2), 327–334. doi:10.1038/sj.bjc.6605154

Morrissy, A. S., Cavalli, F. M. G., Remke, M., Ramaswamy, V., Shih, D. J. H., Holgado, B. L., et al. (2017). Spatial Heterogeneity in Medulloblastoma. *Nat. Genet.* 49, 780–788. doi:10.1038/ng.3838

Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., et al. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* 3, 385–394. doi:10.1016/j.cels.2016.09.002

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., et al. (2011). Tumour Evolution Inferred by Single-Cell Sequencing. *Nature* 472 (7341), 90–94. doi:10.1038/nature09807

Nekrasova, O., and Green, K. J. (2013). Desmosome Assembly and Dynamics. *Trends Cell Biol.* 23, 537–546. doi:10.1016/j.tcb.2013.06.004

Nieto, M. A., Huang, R. Y.-J., Jackson, R. A., and Thiery, J. P. (2016). Emt: 2016. *Cell* 166, 21–45. doi:10.1016/j.cell.2016.06.028

Noren, N. K., Foos, G., Hauser, C. A., and Pasquale, E. B. (2006). The EphB4 Receptor Suppresses Breast Cancer Cell Tumorigenicity through an Abl-Crk Pathway. *Nat. Cell Biol* 8, 815–825. doi:10.1038/ncb1438

Pasquale, E. B. (2008). Eph-Ephrin Bidirectional Signaling in Physiology and Disease. *Cell* 133, 38–52. doi:10.1016/j.cell.2008.03.011

Pastushenko, I., Brisebarre, A., Sifrim, A., Fioramonti, M., Revenco, T., Boumahdi, S., et al. (2018). Identification of the Tumour Transition States Occurring during EMT. *Nature* 556, 463–468. doi:10.1038/s41586-018-0040-3

Pastushenko, I., and Blanpain, C. (2019). EMT Transition States during Tumor Progression and Metastasis. *Trends Cell Biol.* 29, 212–226. doi:10.1016/j.tcb.2018.12.001

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma. *Science* 344, 1396–1401. doi:10.1126/science.1254257

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* 171, 1611–1624. doi:10.1016/j.cell.2017.10.044

Revenco, T., Nicodème, A., Pastushenko, I., Sznurkowska, M. K., Latil, M., Sotiropoulou, P. A., et al. (2019). Context Dependency of Epithelial-To-Mesenchymal Transition for Metastasis. *Cell Rep.* 29, 1458–1468. doi:10.1016/j.celrep.2019.09.081

Sacchetti, A., Teeuwssen, M., Verhagen, M., Joosten, R., Xu, T., Stabile, R., et al. (2021). Phenotypic Plasticity Underlies Local Invasion and Distant Metastasis in colon Cancer. *eLife* 10, e61461. doi:10.7554/elife.61461

Sottoriva, A., Spiteri, I., Piccirillo, S. G. M., Touloumis, A., Collins, V. P., Marioni, J. C., et al. (2013). Intratumor Heterogeneity in Human Glioblastoma Reflects Cancer Evolutionary Dynamics. *Proc. Natl. Acad. Sci.* 110, 4009–4014. doi:10.1073/pnas.1219747110

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

Steinle, J. J., Meininger, C. J., Forough, R., Wu, G., Wu, M. H., and Granger, H. J. (2002). Eph B4 Receptor Signaling Mediates Endothelial Cell Migration and Proliferation via the Phosphatidylinositol 3-Kinase Pathway. *J. Biol. Chem.* 277, 43830–43835. doi:10.1074/jbc.m207221200

Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., et al. (2021). Highly Sensitive Spatial Transcriptomics at Near-Cellular Resolution with Slide-seqV2. *Nat. Biotechnol.* 39, 313–319. doi:10.1038/s41587-020-0739-1

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq. *Science* 352, 189–196. doi:10.1126/science.aad0501

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., et al. (2019). High-definition Spatial Transcriptomics for *In Situ* Tissue Profiling. *Nat. Methods* 16, 987–990. doi:10.1038/s41592-019-0548-y

Vuoriluoto, K., Haugen, H., Kiviluoto, S., Mpindi, J.-P., Nevo, J., Gjerdrum, C., et al. (2011). Vimentin Regulates EMT Induction by Slug and Oncogenic H-Ras and Migration by Governing Axl Expression in Breast Cancer. *Oncogene* 30, 1436–1448. doi:10.1038/onc.2010.509

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation (N Y)* 2, 100141. doi:10.1016/j.xinn.2021.100141

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., et al. (2010). Distant Metastasis Occurs Late during the Genetic Evolution of Pancreatic Cancer. *Nature* 467 (7319), 1114–1117. doi:10.1038/nature09515

Yang, A. D., Camp, E. R., Fan, F., Shen, L., Gray, M. J., Liu, W., et al. (2006). Vascular Endothelial Growth Factor Receptor-1 Activation Mediates Epithelial to Mesenchymal Transition in Human Pancreatic Carcinoma Cells. *Cancer Res.* 66, 46–51. doi:10.1158/0008-5472.can-05-3086

Yang, J., Antin, P., Berx, G., Blanpain, C., Brabletz, T., Bronner, M., et al. (2020). Guidelines and Definitions for Research on Epithelial-Mesenchymal Transition. *Nat. Rev. Mol. Cell Biol* 21, 341–352. doi:10.1038/s41580-020-0237-9

Yang, N.-Y., Pasquale, E. B., Owen, L. B., and Ethell, I. M. (2006). The EphB4 Receptor-Tyrosine Kinase Promotes the Migration of Melanoma Cells through Rho-Mediated Actin Cytoskeleton Reorganization. *J. Biol. Chem.* 281, 32574–32586. doi:10.1074/jbc.m604338200

You, L., Su, P.-R., Betjes, M., Rad, R. G., Beerens, C., van Oosten, E., et al. (2021). Functional Single Cell Selection and Annotated Profiling of Dynamically Changing Cancer Cells. *Nat. Biomed. Eng.* In Press. doi:10.1101/2021.10.12.464054

Zhang, Y., Xi, S., Chen, J., Zhou, D., Gao, H., Zhou, Z., et al. (2017). Overexpression of LAMC1 Predicts Poor Prognosis and Enhances Tumor Cell Invasion and Migration in Hepatocellular Carcinoma. *J. Cancer* 8, 2992–3000. doi:10.7150/jca.21038

Zhao, M., Kong, L., Liu, Y., and Qu, H. (2015). dbEMT: an Epithelial-Mesenchymal Transition Associated Gene Resource. *Sci. Rep.* 5, 11459. doi:10.1038/srep11459

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.

METHODS
published: 23 February 2022
doi: 10.3389/fbioe.2022.842019

scMelody: An Enhanced Consensus-Based Clustering Model for Single-Cell Methylation Data by Reconstructing Cell-to-Cell Similarity

Qi Tian[1], Jianxiao Zou[1,2,3], Jianxiong Tang[1], Liang Liang[4], Xiaohong Cao[5] and Shicai Fan[1,2,3]*

[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China, [2]Intelligent Terminal Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, Chengdu, China, [3]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China, [4]Cancer Center, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China, [5]Department of Geriatric Endocrinology, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China

Single-cell DNA methylation sequencing technology has brought new perspectives to investigate epigenetic heterogeneity, supporting a need for computational methods to cluster cells based on single-cell methylation profiles. Although several methods have been developed, most of them cluster cells based on single (dis)similarity measures, failing to capture complete cell heterogeneity and resulting in locally optimal solutions. Here, we present scMelody, which utilizes an enhanced consensus-based clustering model to reconstruct cell-to-cell methylation similarity patterns and identifies cell subpopulations with the leveraged information from multiple basic similarity measures. Besides, benefitted from the reconstructed cell-to-cell similarity measure, scMelody could conveniently leverage the clustering validation criteria to determine the optimal number of clusters. Assessments on distinct real datasets showed that scMelody accurately recapitulated methylation subpopulations and outperformed existing methods in terms of both cluster partitions and the number of clusters. Moreover, when benchmarking the clustering stability of scMelody on a variety of synthetic datasets, it achieved significant clustering performance gains over existing methods and robustly maintained its clustering accuracy over a wide range of number of cells, number of clusters and CpG dropout proportions. Finally, the real case studies demonstrated the capability of scMelody to assess known cell types and uncover novel cell clusters.

Keywords: single-cell, DNA methylation, epigenetic heterogeneity, consensus-based clustering, cell-to-cell similarity

Edited by:
Jiaofang Shao,
Nanjing Medical University, China

Reviewed by:
Honglei Ren,
University of California, Irvine,
United States
Jian Wu,
Nanjing Medical University, China

****Correspondence:***
Shicai Fan
shicaifan@uestc.edu.cn

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *23 December 2021*
Accepted: *24 January 2022*
Published: *23 February 2022*

Citation:
Tian Q, Zou J, Tang J, Liang L, Cao X and Fan S (2022) scMelody: An Enhanced Consensus-Based Clustering Model for Single-Cell Methylation Data by Reconstructing Cell-to-Cell Similarity. Front. Bioeng. Biotechnol. 10:842019. doi: 10.3389/fbioe.2022.842019

1 INTRODUCTION

As a heritable covalent chemical modification, DNA methylation is closely correlated with cell growth, differentiation, and transformation, which plays decisive roles in diseases and tumorigenesis (Aran and Hellman, 2013; Oakes et al., 2016; Koch et al., 2018). Technological advances have enabled DNA methylation assay at single-nucleotide resolution through high-throughput sequencing (Cokus et al., 2008; Sandoval et al., 2011; Krueger et al., 2012), thus paving the way for quantifying the methylation landscapes across different tissues and individuals. However, bulk protocols typically require thousands to millions of cells per experiment, making it difficult to study rare cell populations or explore the intercellular epigenetic heterogeneity (Schwartzman and Tanay, 2015). With increasing evidence of epigenetic heterogeneity in phenotypically similar cells (Angermueller et al., 2016; Hui et al., 2018), the single-cell methylation sequencing (scME-seq) protocols have demonstrated their capability for the deconvolution of mixed cell populations, such as scBS (Smallwood et al., 2014), scRRBS (Guo et al., 2013), and scWGBS (Farlik et al., 2015). Besides, the parallel single-cell sequencing protocols, like scM&T-seq (Angermueller et al., 2016), scTrio-seq (Hou et al., 2016), and scNOMe-seq (Pott, 2017), have brought new sights into understanding the regulatory mechanisms of epigenetic modifications on transcriptional variation. Although single-cell RNA sequencing (scRNA-Seq) has been widely used for investigating cell heterogeneity, it mainly informs about highly expressed transcripts while scME-seq enables detecting the methylation status of CpGs across gene and non-gene regions (Luo et al., 2017). Moreover, DNA methylation landscapes are not affected by the environment and can be more stable over the lifespan (Lister et al., 2013; Mo et al., 2015). Therefore, how to uncover cellular heterogeneity based on single-cell methylation data is gaining more attention.

To our knowledge, most existing methods incorporated different (dis)similarity relationships between cells into the distance-based clustering algorithms, such as hierarchical clustering (HC), to generate cell partitions. For instance, Farlik et al. clustered cells based on the average methylation over putative regulatory regions using HC with Euclidean distance and complete linkage (Farlik et al., 2016). Besides, a sliding window approach (Smallwood et al., 2014) was proposed to estimate CpG methylation rates and then cells were clustered based on the estimated methylation levels of most variable CpGs (Smallwood et al., 2014) or gene bodies (Angermueller et al., 2016) using Euclidean distance and HC. In addition to the Euclidean distance, the Pearson correlation coefficient was also used to measure cell-to-cell methylation distance and has been combined with the HC algorithm to generate cell partitions based on the site-level (Hou et al., 2016) or region-level (Pott, 2017) methylation. Hui et al. developed PDclust to identify cell types using a pairwise dissimilarity (PD) measure and HC, where the PD value was defined as the average of the absolute difference in methylation status at overlapping CpGs between cell pairs (Hui et al., 2018). Despite the considerable diversity in these clustering methods, different (dis)similarity measures could have a significant effect on the quality of the clustering results in distance-based clustering algorithms and no single measure was appropriate for all situations (Yona et al., 2006; Khalifa et al., 2009; Shirkhorshidi et al., 2015). Moreover, only PDclust was verified across different datasets while the clustering performances of other distance measures on different datasets have not been fully evaluated. Recently, a probabilistic hierarchical mixture model Epiclomal was proposed to cluster cells through pooling information across cells and neighboring CpGs (de Souza et al., 2020). But Epiclomal required several non-probabilistic methods for clustering initialization and failed to consistently achieve clustering performance gains than single-distance-based methods on some real datasets. Additionally, Kapourani et al. (2021) proposed the Bayesian models for single-cell methylation data analysis but focused on their evaluation on missing data imputation (Kapourani and Sanguinetti, 2019) and identifying variable features. In summary, additional clustering methodologies that are universal to different kinds of single-cell methylation datasets are still urgently needed.

Recent advancements in ensemble clustering (Ghaemi et al., 2009; Vega-Pons and Ruiz-Shulcloper, 2011; Boongoen and Iam-On, 2018) have demonstrated that integrating various basic cell partitions in a consensus matrix is effective to generate improved clustering solutions (Kiselev et al., 2017; Zhu et al., 2020; Cui et al., 2021; Wang et al., 2021). The rationale for this idea is to construct a cell-to-cell pairwise similarity matrix based on the diverse basic clustering results through a cluster-based similarity partitioning algorithm (CSPA) (Strehl and Ghosh, 2002), with each value in the matrix representing the probability of the occurrence of cell pairs in the same cluster. Then the resulting ensemble cell clusters can be yielded according to the consensus matrix with typical clustering algorithms, such as HC. Since how to accurately capture intercellular methylation (dis)similarity relationships is significant for clustering cells, combining information from multiple (dis)similarity measures to reconstruct the cell-to-cell similarity with the consensus-based clustering strategy becomes a promising alternative. However, the traditional consensus strategy only integrated the information of basic clustering assignments (Golalipour et al., 2021; Zhang, 2021), which might be not sufficiently informative to reconstruct the cell-to-cell similarity as the inherent distance relationships within the subpopulation were ignored. Moreover, when calculating the consensus matrix, the basic clustering partitions could be highly correlated or differ significantly and their ability to distinguish cells was different, requiring an extra strategy to balance the diversity and separability of the basic clustering partitions. Although many weighting strategies based on various clustering validation indices have been proposed to construct a more accurate consensus matrix (Vega-Pons et al., 2008; Vega-Pons et al., 2011; Ünlü and Xanthopoulos, 2019; Zhu et al., 2020), they did not take into account the diversity and separability of basic cluster partitions simultaneously.

Here, we propose scMelody, an enhanced consensus-based clustering model for single-cell methylation data analysis by reconstructing cell-to-cell pairwise similarity. By introducing a regularization process and a dual weighting strategy, scMelody improves the construction of the consensus matrix which contributes to a novel cell-to-cell similarity measure for clustering cells. Compared to the single (dis)similarity measures, the reconstructed cell-to-cell similarity measure combines the multiple inherent distance relationships of cells and the clustering information of basic cell clusters, so as to improve the accuracy of identifying cell subpopulations. As an additional benefit, scMelody can conveniently leverage the internal clustering validation criterion to determine the optimal number of clusters based on the reconstructed pairwise similarity patterns. Extensive assessments on both real datasets and synthetic datasets showed that scMelody achieved

TABLE 1 | Overview of the seven real single-cell methylation datasets.

Datasets	Sequencing	# GEO accession	# Cells	# Clusters
Smallwood	ScBS	GSE56879	32	2
Farlik2015	scWGBS	GSE65196	69	4
Hou	scTrio-seq	GSE65364	31	3
Pott	scNOMe-seq	GSE83882	23	2
Farlik2016	scWGBS	GSE87197	122	6
Luo-human	snmC-seq	GSE97179	2740	21
Luo-mouse	snmC-seq	GSE97179	3377	16

the most advanced performance over previous methods in clustering single-cell methylation data.

2 MATERIALS AND METHODS

2.1 Datasets and Pre-Processing

We first retrieved seven real single-cell methylation datasets in which cell types were known a priori or were validated in the respective study to benchmark the performance of the clustering algorithms. These distinct single-cell methylation datasets were generated by various sequencing techniques and came from Smallwood et al. (2014), Farlik et al. (2015), Hou et al. (2016), Pott (2017), Farlik et al. (2016) and Luo et al. (2017). The Smallwood dataset was made up of mouse embryonic stem cells (ESCs), where the cells were cultured in a regular serum medium and 2i medium to introduce differential methylation. Note that there were two outlier cells from the serum condition that were demonstrated to be more similar to the 2i ESCs. The Falik2015 dataset consisted of K562 cells and HL60 cells, which were either treated with extra drugs or not, leading to 4 different cell subpopulations. The Hou dataset consisted of the cells were from a human hepatocellular carcinoma (HCC) tissue sample and a human hepatoblastoma-derived cell line (HepG2). There were two subpopulations in HCC cells, where the authors integrated gene expression, copy number changes and DNA methylation to support their findings. The Pott dataset consisted of GM12878 cells and K562 cells, which were grown in different culture mediums. The Farlik2016 dataset contained several different types of human hematopoietic cells, including hematopoietic stem cells (HSC), multipotent progenitors (MPP), common lymphoid progenitor (CLP), common myeloid progenitor (CMP), immature multi-lymphoid progenitor (MLP0), and granulocyte-macrophage progenitor (GMP). The Luo dataset was relatively large, which consisted of two different parts, including 2740 human neurons (Luo-human) and 3,377 mouse neurons (Luo-mouse). According to the original experiment, both the human and mouse neurons were very heterogeneous, where there were 21 subclusters identified in human neurons and 16 subclusters identified in mouse neurons. The overview of these real datasets is summarized in **Table 1**, including the number of cells and the number of clusters for each dataset. Moreover, in addition to the aforementioned datasets for the standard validation, we also retrieved one of the largest publicly available datasets, which assayed 28077 inhibitory neurons from different regions of the mouse brain and presented strong cellular heterogeneity (Liu et al., 2021). We focused on the evaluation of the ability of scMelody to identify novel cell clusters under complex cell composition contexts on this large dataset.

To faithfully simulate methylation data that resemble scME-seq for evaluating the clustering stability and scalability of scMelody, we also generated synthetic datasets with various initial settings using the sub-sampling strategy proposed by Kapourani and Sanguinetti (2019). To retain the structure of missing data observed in sequencing experiments, this strategy generated the pseudo-single cells by sampling the raw FASTQ files of the bulk data. We collected the bulk RRBS data (GEO accession: GSE27584) of 10 cell lines (**Supplementary Table S1**) from the ENCODE dataset (Wang et al., 2012) and the pseudo-single cells were produced by randomly keeping 10% of the mapped reads from the bulk experiment. Then, we generated the synthetic datasets with different initial settings: (1) the number of pseudo-single cells ($N = 50, 100, 200, 300, 400, 500, 600, 800, 1000$); (2) the number of predefined clusters (C = 2, 3, 4, 5, 6, 7, 8, 9, 10); (3) the dropout CpG proportions (η = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). Note that the number of predefined clusters was achieved by combining the cells sampled from different cell lines and we sampled the equal numbers of cells in each cell line. The dropout CpG proportion simulated the data with different sparsity by randomly eliminating a certain proportion of CpG sites in pseudo-single cells, where the higher the dropout proportions represented the higher the degree of data sparsity and the greater difficulty of clustering. In comparative studies, we varied one parameter and kept the others fixed. Unless otherwise specified, the fixed parameters were: number of pseudo-single cells 400, number of predefined clusters 6 and the CpG dropout proportion 0.5. For each setting, we generated 50 input datasets to evaluate the clustering performance.

For the retrieved real single-cell methylation datasets, most of the CpG loci assayed exhibited binary methylation status (methylated or unmethylated). Specifically, the CpGs detected by snmC-seq only had methylated or unmethylated status and the CpGs detected by other sequencing techniques predominantly presented either hypermethylation or hypomethylation (**Supplementary Figure S1**). Considering the bimodal distribution of methylation levels, the CpGs exhibiting partially methylated calls (≥.5) were assigned a value of 1 (methylation) or a value of 0 (unmethylation) otherwise (<.5). Similarly, for the synthetic datasets generated from the RRBS bulk data, the binary methylation status could be obtained by using a threshold of .5 (values no less than .5 were binarized to 1 otherwise to 0).

2.2 scMelody Clustering Algorithm

Considering the sparse coverage of scME-seq technology, scMelody leverages all overlapping CpGs between cell pairs to evaluate cell-to-cell similarity patterns. Specifically, scMelody takes files with binary CpG methylation calls across the genome from individual cells as input. To capture different methylation similarity patterns between cell pairs, scMelody utilizes three correlation-based measures, including Cosine, Hamming and Pearson correlation coefficient, which have been reported to be effective for quantifying the similarity relationships of binary data (Haranczyk and Holliday, 2008). Given a series of single-cell methylation data files X_i (i = 1,2 . . ., n; n denotes the number of target cells), the Cosine similarity of cell pairs (X_i, X_j) can be calculated as follows:

$$S_1\left(X_i, X_j\right) = \frac{\sum_{t=1}^{m} X_{it} X_{jt}}{\sqrt{\sum_{t=1}^{m} (X_{it})^2 \sum_{t=1}^{m} \left(X_{jt}\right)^2}}$$

where m represents the number of overlapping CpGs shared by cell pairs (X_i, X_j) and t denotes t -th overlapping CpG between each cell pair (X_i, X_j). For any two cells, the more similar the global methylation landscape is, the larger the Cosine correlation coefficient is; and $S_1(X_i, X_j)$ ranges from 0 to 1. Next, scMelody calculates the Hamming similarity for each cell pair (X_i, X_j):

$$S_2\left(X_i, X_j\right) = \frac{\sum_{t=1}^{m} I\left(X_{it} = X_{jt}\right)}{m}$$

where the indicator function $I(.)$ returns 1 if its argument is true. This can be described as calculating the proportion of CpGs with concordant methylation status between cell pairs, which ranges from 0 to 1. Finally, the Pearson similarity is calculated as follows:

$$S_3\left(X_i, X_j\right) = \frac{\sum_{t=1}^{m}\left(X_{it} - \overline{X_i}\right)\left(X_{jt} - \bar{X}_j\right)}{\sqrt{\sum_{t=1}^{m}\left(X_{it} - \overline{X_i}\right)^2}\sqrt{\sum_{t=1}^{m}\left(X_{jt} - \overline{X_j}\right)^2}}$$

where $\overline{X_i}$, $\overline{X_j}$ is the mean of X_i, X_j respectively and the Pearson similarity measures the linear correlation according to the methylation status between the cell pair (X_i, X_j), varying from 0 to 1. With the three basic similarity measures, the inherent methylation similarity relationships of cells can be quantified and the cell-to-cell methylation similarity patterns are captured in the corresponding similarity matrices $\{S_\mu | \mu = 1, 2, 3\}$.

To reconstruct the cell-to-cell methylation similarity with the consensus-based clustering strategy, scMelody implements spectral clustering (von Luxburg, 2007) to generate basic cell partitions according to the methylation similarity matrices. Spectral clustering does not make strong assumptions on the form of the cluster and is effective for clustering sparse data with only similarity relationships between data points. Given a similarity matrix $S = (s_{ij}) \in \mathbb{R}^{n\times n}$, where $s_{ij} \geq 0$ represents the linkage weights between cell i and cell j, spectral clustering partitions the cells into C clusters through solving the following optimization problem:

$$\underset{L \in \mathbb{R}^{n\times C}}{min} < LL^T, I_n - \tilde{S} >, s.t. L^T L = I_C$$

where $\tilde{S} = D^{-1/2} S D^{-1/2}$ and $D = \mathrm{diag}(d_{11}, d_{22}, \ldots, d_{nn})$ is a diagonal matrix with $d_{ii} = \sum_{j=1}^{n} s_{ij}$. Finally, each row of obtained L is treated as a data point in $\mathbb{R}^C$, and is clustered into C groups by k-means. Note that $I_n - \tilde{S}$ is called a normalized graph Laplacian. By implementing spectral clustering on the three similarity matrices $\{S_1, S_2, S_3\}$, we can generate a set of basic cell partitions $\mathbf{\Pi} = \{\pi_\mu | \mu = 1, 2, 3\}$, which can be used as a clustering prior for reconstructing cell-to-cell similarity.

To convert the information of each basic cell partition into the respective cell-to-cell similarity matrix, scMelody constructs a co-occurrence matrix for each basic cluster. In traditional consensus clustering strategy, for each basic clustering assignment π_μ in $\mathbf{\Pi}$, an $n \times n$ binary co-occurrence matrix is constructed, which can be denoted as I_μ:

$$I_\mu\left(X_i, X_j\right) = \begin{cases} 1 & if C(X_i) = C\left(X_j\right) \\ 0 & otherwise \end{cases}$$

where $C(X_i)$ denotes the clustering label of cell X_i, and if the cell pairs (X_i, X_j) are assigned into the same cluster in the μ -th member π_μ, the value of $I_\mu(X_i, X_j)$ is equal to 1, otherwise is 0. The general consensus matrix is obtained by averaging the binary co-occurrence matrices I_μ. However, this may not be sufficiently informative to reconstruct cell-to-cell similarity as the inherent similarity relationships of cells are ignored and the resulting consensus matrix is heavily dependent on the basic cell partitions.

To reconstruct the cell-to-cell similarity patterns that faithfully reflects the methylation difference between cells, scMelody adopts a two-stage strategy to improve the construction of the consensus matrix and the resulting consensus matrix can be used to measure the cell-to-cell pairwise similarity in higher resolution. In the first stage, scMelody redefines the construction of the binary co-occurrence matrix I_μ to produce a more fine-grained co-occurrence matrix I_μ^*. Specifically, scMelody utilizes the basic similarity matrix to regularize the binary co-occurrence matrix I_μ and the new co-occurrence matrix I_μ^* can be expressed as:

$$I_\mu^* = I_\mu \odot S_\mu$$

where $\odot$ denotes the Hadamard product and each value in I_μ^* can be calculated as $I_\mu^*(X_i, X_j) = I_\mu(X_i, X_j) \times S_\mu(X_i, X_j)$. In this way, the new matrix I_μ^* measures the co-occurrence of cell pairs belonging to the same cluster in higher resolution. Compared to I_μ, I_μ^* refines the similarity of cells within the clusters, while preserving the differences between cells belonging to different clusters. In the second stage, scMelody adaptively assigns weights to different I_μ^* based on the diversity and separability of the basic cell partitions with a dual weighting strategy. Firstly, existing studies have underlined the importance of diversity in basic clustering partitions to enhance the performance of ensemble solutions (Kuncheva and Hadjitodorov, 2004; Hadjitodorov et al., 2006; Fern et al., 2008), thus scMelody proposes a weighting criterion to assess the diversity of basic cell partitions based on NMI (Vinh et al., 2010), where NMI utilizes mutual information to measure the agreement of the two clustering assignments. Suppose each basic cell partition $\pi_\mu = \{C_1^\mu, C_2^\mu, \ldots C_k^\mu, \ldots, C_{K\mu}^\mu\}$, C_k^μ is a cluster of π_μ and $K\mu$ denotes the number of the clusters of π_μ. To punish the basic cell partition that contributes little to the diversity, the weight for basic cell partition π_μ can be formularized as follows:

$$w_\mu^{div} = \frac{exp\left(-\frac{1}{r-1}\sum_{\nu=1,\nu\neq\mu}^{r} NMI\left(\pi_\mu, \pi_\nu\right)\right)}{\sum_{\mu=1}^{r} exp\left(-\frac{1}{r-1}\sum_{\nu=1,\nu\neq\mu}^{r} NMI\left(\pi_\mu, \pi_\nu\right)\right)}$$

$$NMI(\pi_\mu, \pi_\nu) = \frac{2 \times \sum_{k,l} p_{kl} log \frac{p_{kl}}{p_k \times p_l}}{-\sum_k p_k log p_k - \sum_l p_l log p_l}$$

where $r = 3$ represents the number of basic cell partitions. Besides, $p_k = n_k/n$, $p_l = n_l/n$ and $p_{kl} = n_{kl}/n$, where n_k, n_l represents the number of cells in the k -th and l -th cluster of the basic cell partition π_μ, π_ν respectively, and n_{kl} is the number of cells shared by cluster k and cluster l. NMI score ranges from 0 to 1, with higher NMI score representing more consistent basic cell partitions and $\frac{1}{r-1}\sum_{\nu=1,\nu\neq\mu}^{r} NMI(\pi_\mu, \pi_\nu)$ measures the overall consistency between the basic cell partition π_μ and others, with higher values representing less contribution to the diversity. Note that $0 < w_\mu^{div} < 1$ and $\sum_\mu w_\mu^{div} = 1$. Then, to assess the separability of basic cell partitions, scMelody considers the silhouette coefficient (Rousseeuw, 1987), which combines the cohesion and separation of clusters to assess the clustering performance when the ground truth labels are not known. Given a basic cell clustering assignment $\pi_\mu = \{C_1^\mu, C_2^\mu, \ldots C_k^\mu, \ldots, C_{K\mu}^\mu\}$, the weight defined by the separability can be obtained as follows:

$$w_\mu^{sep} = \frac{exp(SI(\pi_\mu))}{\sum_{\mu=1}^{r} exp(SI(\pi_\mu))}$$

$$SI(\pi_\mu) = \frac{1}{K\mu}\sum_k \left\{ \frac{1}{n_k} \sum_{X_i \in C_k^\mu} \frac{b(X_i) - a(X_i)}{\max[b(X_i), a(X_i)]} \right\}$$

$$a(X_i) = \frac{1}{n_k - 1}\sum_{X_j \in C_k^\mu, X_j \neq X_i} \boldsymbol{K}(S_\mu(X_i, X_j))$$

$$b(X_i) = min_{l, l\neq k}\left\{ \frac{1}{n_l}\sum_{X_j \in C_l^\mu} \boldsymbol{K}(S_\mu(X_i, X_j)) \right\}$$

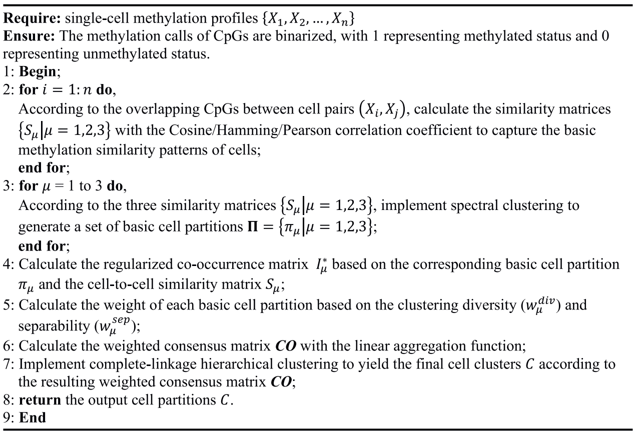
where $a(X_i)$ denotes the average distance between cell X_i and all other cells in the same cluster C_k^μ while $b(X_i)$ denotes the average distance between cell X_i and all other cells in the next nearest cluster C_l^μ. Here, $\boldsymbol{K}(.)$ is a kernel function that converts the similarity measure $S_\mu(X_i, X_j)$ to the respective distance measure $1 - S_\mu(X_i, X_j)$ as the original value of the basic cell-to-cell similarity measure varies from 0 to 1. $SI(\pi_\mu)$ ranges from -1 to 1, with a higher value indicating that the intra-class distance is small while the inter-class distance is large thus the cells are well-clustered. Note that we also have $0 < w_\mu^{sep} < 1$ and $\sum_\mu w_\mu^{sep} = 1$, with higher w_μ^{sep} indicating higher separability for basic cell partition π_μ. In this way, scMelody achieves the assessment of weights based on the diversity and separability of the basic cell partitions. Combining with the regularized co-occurrence matrix I_μ^*, the resulting weighted consensus matrix $\boldsymbol{CO}$ can be constructed through a linear aggregation function, which can be expressed as:

$$CO(X_i, X_j) = \boldsymbol{f}(w, I^*) = 0.5 * \left(\sum_\mu w_\mu^{div} I_\mu^* + \sum_\mu w_\mu^{sep} I_\mu^* \right)$$

where 0.5 is used as a scaling coefficient, restricting the value of cell-to-cell pairwise similarity in the weighted consensus matrix $\boldsymbol{CO}$ varying from 0~1. Each value $CO(X_i, X_j)$ in the resulting weighted consensus matrix is a reconstructed similarity measure of each cell pair (X_i, X_j), which measures the methylation similarity relationships between cells in higher resolution.

Finally, the weighted consensus matrix $\boldsymbol{CO}$ is clustered using the complete-linkage HC algorithm to yield the resulting cell partitions. The overall scMelody clustering framework is shown in **Figure 1**, and the pseudo code flow is available in **Algorithm 1**.

Algorithm 1: scMelody



```
Require: single-cell methylation profiles $\{X_1, X_2, \ldots, X_n\}$
Ensure: The methylation calls of CpGs are binarized, with 1 representing methylated status and 0
representing unmethylated status.
1: Begin;
2: for $i = 1:n$ do,
    According to the overlapping CpGs between cell pairs $(X_i, X_j)$, calculate the similarity matrices
    $\{S_\mu | \mu = 1,2,3\}$ with the Cosine/Hamming/Pearson correlation coefficient to capture the basic
    methylation similarity patterns of cells;
    end for;
3: for $\mu = 1$ to 3 do,
    According to the three similarity matrices $\{S_\mu | \mu = 1,2,3\}$, implement spectral clustering to
    generate a set of basic cell partitions $\mathbf{\Pi} = \{\pi_\mu | \mu = 1,2,3\}$;
    end for;
4: Calculate the regularized co-occurrence matrix $I_\mu^*$ based on the corresponding basic cell partition
    $\pi_\mu$ and the cell-to-cell similarity matrix $S_\mu$;
5: Calculate the weight of each basic cell partition based on the clustering diversity ($w_\mu^{div}$) and
    separability ($w_\mu^{sep}$);
6: Calculate the weighted consensus matrix $\boldsymbol{CO}$ with the linear aggregation function;
7: Implement complete-linkage hierarchical clustering to yield the final cell clusters $C$ according to
    the resulting weighted consensus matrix $\boldsymbol{CO}$;
8: return the output cell partitions $C$.
9: End
```

2.3 Determine the Optimal Number of Clusters

Both the spectral clustering and HC algorithms need to specify the number of clusters in advance to generate the cluster assignments. Here, we integrate basic similarity measures of cells to propose a robust strategy to determine the optimal number of clusters based on the silhouette coefficient criterion. Let $k = \{2, \ldots, K_{max}\}$, where K_{max} denotes the possible maximum number of clusters, we first run the spectral clustering varying k (k denotes the input number of clusters for spectral clustering) from 2 to K_{max}. Let π_k represents the corresponding cell partition when the input number of clusters equaling k. For the three different similarity measures, we can get three different cell partitions at each value of k. Then, we calculate the silhouette coefficient for each similarity measure at each k and select the best k_{sp} as the optimal number of spectral clustering which is given by:

$$k_{sp} = argmax\sum_{\mu=1}^{r} (SI(\pi_k)|k)$$

where $r = 3$ represents the number of spectral clustering partitions and $(SI(\pi_k)|k)$ represents silhouette coefficient of the corresponding spectral clustering partition based on similarity measure μ at each k. k_{sp} is selected as the optimal number for spectral clustering when the sum of the corresponding silhouette coefficients generated from the three basic similarity measures reaches maximum. Then, to generate the final cell partitions, the reconstructed similarity matrix $\boldsymbol{CO}$ is clustered using the complete-linkage HC algorithm. We cut the hierarchical tree at k_{opt} clusters which can be expressed as:

$$k_{opt} = argmax(SI(\pi_k)|k)$$

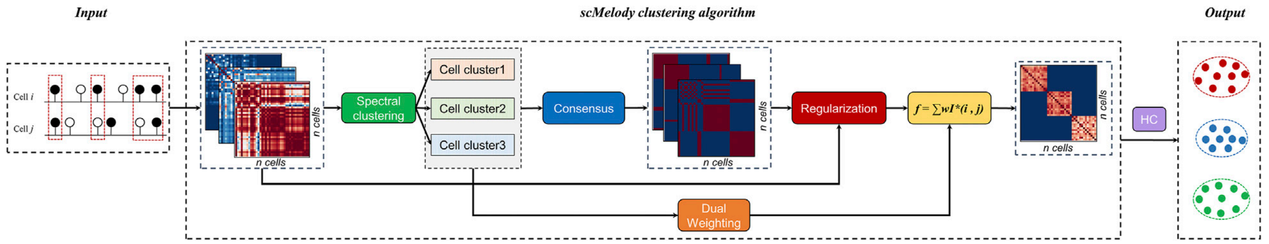


FIGURE 1 | Illustrative flowchart of scMelody. scMelody first utilizes three correlation-based measures to capture cell-to-cell methylation similarity patterns, including Cosine, Hamming and Pearson. The basic cell clusters are generated by spectral clustering according to the similarity patterns. Then, scMelody leverages an enhanced consensus-based clustering model to reconstruct the cell-to-cell similarity by integrating the basic cell clusters and similarity patterns. The resulting cell cluster is generated by performing the complete-linkage hierarchical clustering according to the reconstructed cell-to-cell similarity matrix.

where k_{opt} is the optimal number of the resulting cell partitions and can be obtained when the silhouette coefficient generated from the reconstructed similarity measures reaches maximum.

2.4 Model Comparison

To evaluate the clustering performance of scMelody, we performed intensive comparative studies with previously published methods, which were described as follows:

SW + HC (Smallwood et al., 2014): The sliding window (SW) approach first estimated the sample-specific methylation rates of the genome-wide CpGs in a single cell based on a binomial distribution. To increase the coverage across cells, a sliding window of 3 kb in size and 600 bp in step size was used to subdivide the genome. Then the cell-to-cell methylation variances were evaluated using the estimated sample-specific methylation rates. The cell partitions were generated by the complete-linkage hierarchical clustering.

PearsonHC (Hou et al., 2016): This approach utilized the Pearson correlation coefficient to measure cell-to-cell methylation similarity based on the genome-wide overlapping CpGs of cell pairs. This measure was identical to the Pearson similarity metric used in scMelody. The complete-linkage HC was implemented to generate the cell clusters.

PDclust (Hui et al., 2018): PDclust depended on a measure of CpG methylation pairwise dissimilarity (PD), which was defined as the proportion of the overlapping CpGs with discordant methylation status between each pair of cells. The cell partitions were generated by calculating Euclidean distances between each pair of cells based on their PD values using the Ward-linkage HC. Note that the PD value used in PDclust is different from the Hamming similarity measure in scMelody, as the Hamming similarity measure quantified the methylation similarity of cell pairs and the basic cell partitions were obtained based entirely on Hamming similarity without calculating the Euclidean distances of the measure.

Epiclomal (de Souza et al., 2020): Epiclomal was a probabilistic clustering method arising from a hierarchical mixture model which performed better than single-distance-based methods on several datasets. There were two major variants for Epiclomal, including EpiclomalBasic (EpiclomalB) and EpiclomalRegion (EpiclomalR). EpiclomalB considered the methylation status of all CpGs while EpiclomalR focused on the methylation levels across genomic functional regions such as CGIs, leading to better interpretation of the expected cellular heterogeneity on real datasets. Thus, the author mainly focused on the clustering performance of EpiclomalR on real datasets. To be fair, we applied the two versions of Epiclomal on the synthetic datasets; while on the real datasets, only EpiclomalR was considered. For EpiclomalR, the clustering assignments were generated from the filtered inputs of 10,000 CpGs, which were based on the functional genomic regions from CGI and TFBS.

2.5 Clustering Performance Metrics

To evaluate the performance of different clustering algorithms, we utilize two popular clustering validation indices, including the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and V-measure (Rosenberg and Hirschberg, 2007). Both the two clustering validation indices measure the agreement between the inferred cell clusters and the true or predefined ones from different perspectives. ARI measures clustering performance by the similarity or matching degree between the prediction target cluster vector and the real cluster vector. Given a set of m cells, the quantitative relationship between the clustering results and the reference labels can be reflected in a contingency table, where each entry indicates the number of objects in common between the prediction and the reference.

$$ARI = \frac{\sum_{ij}\binom{m_{ij}}{2} - \left[\sum_i \binom{\alpha_i}{2} \sum_j \binom{\beta_j}{2}\right] \Big/ \binom{m}{2}}{\frac{1}{2}\left[\sum_i \binom{\alpha_i}{2} + \sum_i \binom{\beta_j}{2}\right] - \left[\sum_i \binom{\alpha_i}{2} \sum_j \binom{\beta_j}{2}\right] \Big/ \binom{m}{2}}$$

Where m_{ij} comes from the contingency table, α_i is the sum of the i^{th} row of the contingency table, β_j is the sum of the j^{th} column of the contingency table and the (.) function denotes a binomial coefficient. The V-measure captures the homogeneity and completeness of a clustering result. To satisfy the homogeneity criterion, each cluster contains only members of a single class. Completeness is satisfied if all those cells that are members of a single group are assigned to a single cluster. The V-measure can

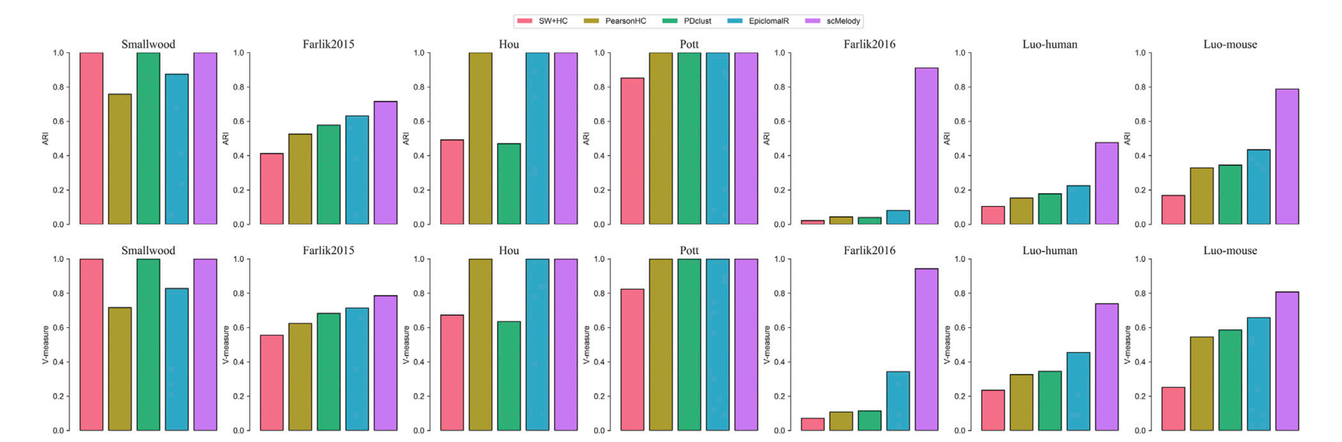


FIGURE 2 | Clustering performance comparison between scMelody and other major published methods on the real datasets. Both ARI and V-measure are employed to assess the similarity between inferred and true cluster labels.

be calculated as the harmonic mean of homogeneity (h) and completeness (c):

$$V = \frac{2hc}{h + c}$$

where the homogeneity $h = 1 - H(C|K)/H(C)$, $H(C|K)$ is the conditional entropy of the classes given the cluster assignments and is given by $H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n}log(\frac{n_{c,k}}{n})$, $H(C)$ is the entropy of the classes and is given by $H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n}log(\frac{n_c}{n})$, with n the number of cells, n_c and n_k the number of cells respectively belonging to true class c and cluster k, and $n_{c,k}$ the number of cells from true class c assigned to cluster k. The completeness $c = 1 - H(K|C)/H(K)$, which has the analogous formulation as the homogeneity h.

3 RESULTS

3.1 scMelody Outperforms the Existing Methods

We first benchmarked scMelody together with the other published methods on 7 real single-cell methylation datasets, reflecting a wide spectrum of sequencing techniques, data sparsity, number and heterogeneity of single cells examined. **Figure 2** showed the clustering performance of these methods across the datasets, which clearly indicated that scMelody outperformed other methods by achieving the same or higher ARI and V-measures scores. Specifically, on the three datasets with fewer cells or clusters, including Smallwood, Hou and Pott, scMelody accurately clustered all cells while other methods misclassified one or several cells. On the Farlik2015 dataset, the heterogeneity between the different cell subpopulations (treated or untreated) was subtle, however, scMelody performed better than the competing methods by achieving less misclassification for both K562 and HL60 treated cells. On the Farlik2016 dataset, scMelody achieved significant clustering performance gains than other methods, where the inferred assignments of scMelody showed much higher consistency with the true cell clusters (**Supplementary Figure S2**). On the two relatively large datasets, scMelody was superior to the competing methods by recapitulating the major cell types more accurately and achieved higher ARI and V-measure scores. Moreover, EpiclomalR accurately identified the cell heterogeneity on both Hou and Pott datasets and was slightly inferior to scMelody on Smallwood and Farlik2015 datasets while was significantly inferior to scMelody on Farlik 2016, Luo-human and Luo-mouse datasets. The clustering performances of the three single-distance-based methods varied a lot across different datasets. On the simple datasets with fewer numbers of cells or clusters (like Smallwood and Pott), they could accurately identify the cell heterogeneity and achieved close ARI or V-measure scores compared to scMeldoy and EpiclomalR; however, their clustering performance decayed rapidly on complex datasets with increasing numbers of cells or clusters (like Farlik2016 and Luo-human). Additionally, we also observed that even the three single-distance-based methods achieved different clustering performances on different datasets and no single measure could always be better than others. **Supplementary Figure S3** summarized the ARI scores and V-measure scores of the benchmarked methods across the real datasets and scMelody showed the highest average ARI and V-measure scores, indicating that our model was universal to different kinds of single-cell methylation datasets.

We further investigated the performance of the benchmarked methods in terms of estimating the number of clusters. Since only EpiclomalR and scMelody provided built-in functions for predicting the number of clusters, we utilized the silhouette coefficient criterion to specify the optimal number of clusters for the three single-distance-based methods. The result showed that all methods accurately estimated the optimal number of clusters on the datasets with the fewer true numbers of clusters, including the Smallwood and Pott datasets (**Table 2**). While on the datasets with stronger cellular heterogeneity, scMelody achieved improved estimations that were closer to the numbers of true clusters, such as accurately predicting the number of clusters on the Farlik2016 and Luo-mouse datasets and achieving smaller prediction errors on the Luo-human

TABLE 2 | The estimated number of clusters on each real dataset.

Datasets	True clusters	SW + HC	PearsonHC	PDclust	EpiclomalR	scMelody
Smallwood	2	2	2	2	2	2
Farlik2015	4	2	2	2	2	2
Hou	3	2	3	3	3	3
Pott	2	2	2	2	2	2
Farlik2016	6	2	3	2	7	6
Luo-human	21	13	14	15	25	18
Luo-mouse	16	10	12	12	15	16

dataset. EpiclomalR provided better prediction performance than the three single-distance-based methods while the three single-distance-based methods generally underestimated the number of clusters. Of note, although scMelody and the three single-distance-based methods all predicted the number of clusters based on the silhouette coefficient criterion, the better prediction performance of scMelody suggested that the reconstructed cell-to-cell similarity enabled a more accurate reflection of the differences between cell subpopulations.

3.2 scMelody Defines a Better Similarity Measure With Improved Clustering Performance

To further illustrate that scMelody could improve the clustering performance by reconstructing cell-to-cell similarity with the proposed enhanced consensus clustering strategy, we further investigated the clustering results generated by different similarity measures. Using the HC as the benchmarked clustering algorithm, the cell partitions were generated from different similarity matrices: 1) The three basic similarity matrices, including Cosine, Hamming and Pearson. 2) Consensus-I, the similarity matrix was the traditional consensus matrix generated by averaging the binary co-occurrence matrices without the regularization process and the weighting process. 3) Consensus-II, the similarity matrix was the consensus matrix generated by averaging the regularized co-occurrence matrices without the weighting process. 4) Consensus-III, the similarity matrix was the consensus matrix generated by weighting the binary co-occurrence matrices without the regularization process. 5) The similarity matrix was the resulting consensus matrix of scMelody. The differences between these similarity measures are summarized in **Table 3**.

TABLE 3 | The differences between the benchmarked similarity measures.

Similarity	Consensus	Regularization	Weighting
Cosine	No	—	—
Hamming	No	—	—
Pearson	No	—	—
Consensus-I	Yes	No	No
Consensus-II	Yes	Yes	No
Consensus-III	Yes	No	Yes
ScMelody	Yes	Yes	Yes

The results showed that the clustering performance varied considerably between different similarity measures (**Figure 3**). Firstly, we observed that the reconstructed cell-to-cell similarity by scMelody could dissect cellular heterogeneity more accurately and robustly, as it achieved better or the same clustering performance than other similarity measures across all the datasets. Secondly, we also observed that the clustering performances of the basic similarity measures varied considerably on different datasets, indicating that they captured methylation differences between cells from different aspects. Thirdly, generally speaking, integrating the information from basic similarity measures could more accurately reflect the true methylation heterogeneity between cells, which was reflected in the improved clustering accuracy of the consensus-based similarity measures than the basic similarity measures on most datasets. However, we also observed that Consensus-I did not consistently improve the clustering performance on all datasets (like the Smallwood, Farlik2015 and Hou datasets) compared to the basic similarity measures, indicating the limitation of the traditional consensus strategy. Moreover, the overall performance of Consensus-I was not as good as Consensus-II or Consensus-III and this suggested that both the regularization and weighting strategy contributed to boosting the clustering performance. In conclusion, the reconstructed similarity measure by scMelody could achieve more significant clustering performance gains than the basic similarity measures across different real datasets.

3.3 Clustering Stability and Scalability of scMelody

After verifying the clustering performance of scMelody on the real datasets, we generated a variety of synthetic datasets to further evaluate its clustering stability, where the clustering complexity could be controlled with different initialization settings. Firstly, we compared the clustering performance of scMelody and other published methods when the number of cells varied over a wide range. The results showed that when we fixed the number of clusters ($C = 6$) and the CpG dropout proportion ($\eta = 0.5$), the clustering performance of all methods improved with the increase of the cell numbers, while scMelody performed better than other methods across all settings of cell numbers (**Figure 4A**). Compared with EpclomalB, EpiclomalR had better average clustering performance when the numbers of cells were small ($N \leq 600$), but EpiclomalB outperformed EpiclomalR when the numbers of cells were relatively large, indicating that using the information

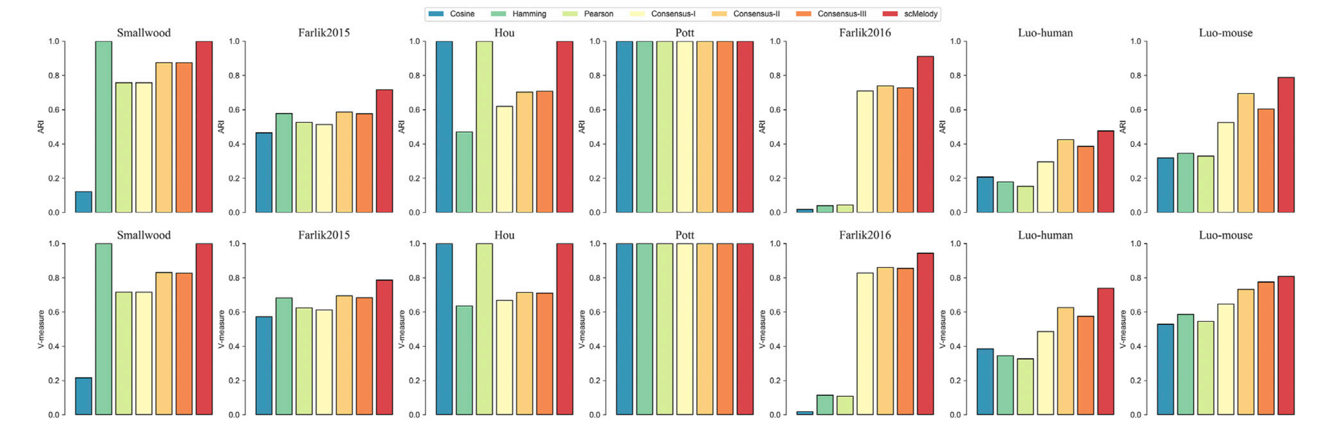


FIGURE 3 | Clustering performance comparison of different similarity measures on the real datasets. These similarity measures include the three basic correlation-based measures and the consensus-based similarity measures. The complete-linkage hierarchical clustering is used as the benchmarked clustering algorithm.

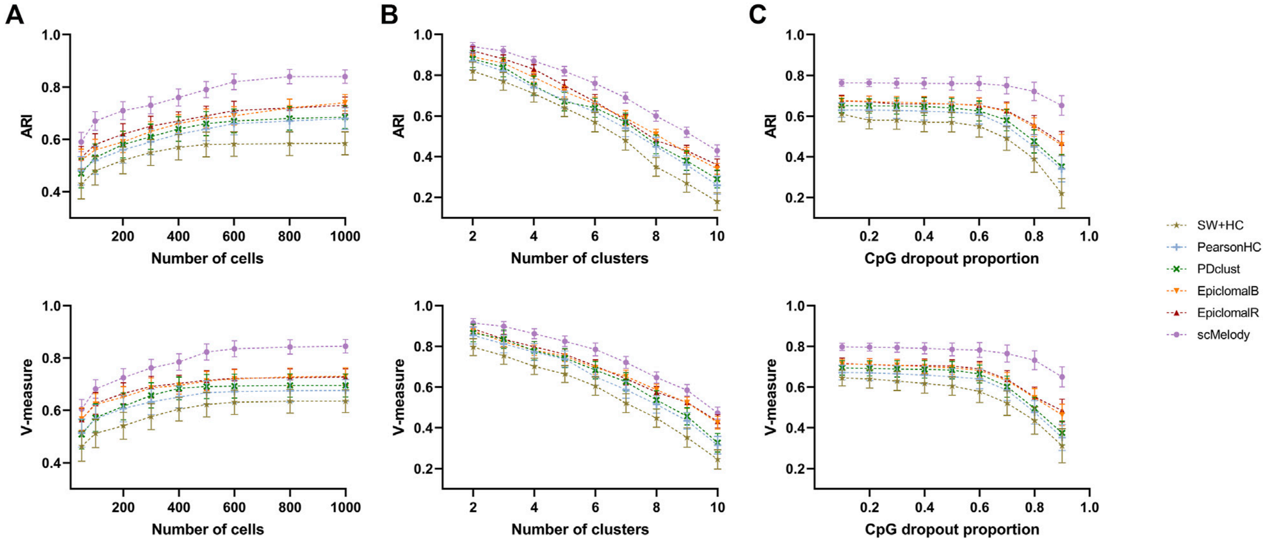


FIGURE 4 | Benchmarking the clustering stability of scMelody and other major published methods on a variety of synthetic datasets. The clustering performance is measured by ARI and V-measure when we vary by: **(A)** number of cells; **(B)** number of clusters; **(C)** CpG dropout proportions. Each setting covers 50 input datasets to evaluate the average clustering performance.

from genome-wide CpGs might better capture cellular heterogeneity than local functional regions when clustering a large number of cells. We also observed that the two correlation-based methods (PearsonHC and PDclust) were better than the method (SW + HC) based on the Euclidean distance. **Figure 4B** showed the clustering performance of the benchmarked methods when varying numbers of clusters (with $N = 600$ and $\eta = 0.5$). When the predefined numbers of clusters were small, the differences in clustering performance among the methods were not significant due to the lower complexity of the clustering task; however, with the increase of the number of clusters, the clustering performance of all methods began to drop while scMelody achieved higher average ARI and V-measure scores than the competing methods. Epiclomal performed better than other single-distance-based clustering methods, while PDclust and PearsonHC were better than SW + HC. Finally, when varying the sparsity of the synthetic datasets by CpG dropout proportions, scMelody achieved better clustering performance under all CpG dropout proportions than the competing methods and could maintain the clustering accuracy across a wide range of dropout proportions ($\eta \leq 0.7$), demonstrating its capability and sensitivity in robustly identifying cell subpopulations (**Figure 4C**).

Furthermore, considering that current single-cell methylation sequencing techniques have already assayed tens to thousands of cells, we also evaluated the runtime of these methods at different

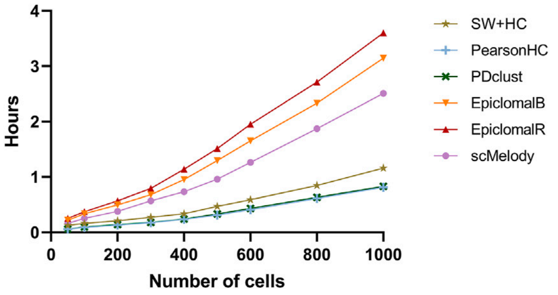


FIGURE 5 | The average runtime of the benchmarked methods on the synthetic datasets with different numbers of cells.

cell numbers. Note that all calculation was performed on a Windows server with an Intel Xeon Platinum 8160 CPU (2.1 GHz) and 32G RAM. **Figure 5** summarized the average time consumption of the benchmarked methods on the synthetic datasets at different numbers of cells. It was obvious that the three single-distance-based methods had lower time consumption than Epiclomal and scMelody, in which SW + HC required more running time than PearsonHC and PDclust. Moreover, scMelody was more computationally efficient compared to EpiclomalB and EpiclomalR while EpiclomalR was more computationally expensive than EpiclomalB. Of note, we found that scMelody spent more than 99% of the running time on calculating the basic cell-to-cell similarity matrices for the input single-cell methylation profiles (**Supplementary Figure S4**) and this was also true for single-distance-based methods, such as PearsonHC and PDclust. Since scMelody was demonstrated to be stable over a wide range of CpG dropout proportions, researchers were recommended to select CpGs from genomic regions of interest to speed up the calculation of the basic similarity matrices in real application scenarios. Besides, considering the varying number of CpGs assayed in real single-cell methylation datasets, **Supplementary Table S2** also showed the runtime of the benchmarked methods on the real datasets and the runtime of scMelody varied within several hours which was practical. To sum up, scMelody could accurately cluster thousands of cells within hours, reaching a balance between the clustering accuracy and the computation efficiency.

3.4 The Reconstructed Similarity Facilitates to the Interpretation of Cell Heterogeneity

To further demonstrate the ability of scMelody to uncover known cell types, we presented two real case studies for the Smallwood and Luo-mouse datasets. Firstly, we investigated whether the cell-to-cell similarity values could visually assess the structures of cell subpopulations, including the reconstructed similarity measure and the three basic similarity measures. **Supplementary Figure S5** showed the heatmaps based on the cell-to-cell pairwise similarity values for the Smallwood dataset. It could be observed that cells with the reconstructed similarity values by scMelody presented a grouping tendency in the diagonal (**Supplementary Figure S5A**), indicating two significant heterogeneous cell populations on this dataset. Combined with the true cell labels, we found that the two major subpopulations were precisely representative of 2i ESCs and serum ESCs. However, even the basic similarity measures also provided accurate clustering results, like Hamming similarity measure, they could not provide the same aggregation tendency in the diagonal as scMelody did (**Supplementary Figures S5B–S5D**). This indicated that the reconstructed cell-to-cell similarity could contribute to the characterization of methylation heterogeneity between cells, which could help researchers intuitively assess the potential cell subpopulations. Secondly, we further investigated the clustering results of the consensus-based similarity measures and focused on the effects of the regularization process and the dual weighting strategy on the output cell clusters. Based on the methylation levels in 100 kb bins across the genome, **Figure 6** showed the t-SNE(van der Maaten and Hinton, 2008) visualization results of the Luo-mouse dataset according to the original cell types and inferred clusters, where the inferred clusters were generated by different consensus clustering strategies, including scMelody, Consensus-II and Consensus-III (**Table 3**). The results indicated that scMelody generated more accurate cell clusters which showed a better agreement with the original cell types. Compared to Consensus-II and Consensus-III, scMelody could more accurately identify the major differences between cell subpopulations and avoid overestimating cellular heterogeneity within the subpopulations. This demonstrated the capability of the enhanced consensus-based clustering model to uncover the cell subpopulations, which could boost the clustering performance by integrating the regularization process and the dual weighting strategy.

3.5 scMelody Uncovers Novel Cell Clusters

To demonstrate the capability of scMelody in identifying novel cell clusters, we presented two case studies. Firstly, according to the annotations from the original experiment of the Farlik2016 dataset, the clustering result of scMelody showed that six cells (denoted as HSC-sub) annotated as HSC were clustered as MPP (**Supplementary Table S3**) while the remaining HSCs (denoted as HSC-raw) were independently grouped together (**Supplementary Figure S6**). To explore the cause of the deviation, we first examined the pairwise methylation similarity of all cells which were annotated as HSC according to their genome-wide methylation status (**Figure 7A**). The result showed that cells denoted as HSC-sub or HSC-raw showed high internal correlations and was much higher than assembling them together (HSC-all), indicating potential heterogeneity among the two subpopulations (HSC-sub and HSC-raw). Then, to provide a biologically meaningful basis for analyzing DNA methylation differences between the HSCs and MPPs, we further aggregated the DNA methylation profiles at the functional genomic region level according to the BLUEPRINT version of the Ensembl Regulatory Build (Zerbino et al., 2015; Adams et al., 2012), including six types of putative regulatory regions. **Figure 7B** showed the t-SNE visualization result of all cells in the Farlik2016 dataset according to their annotated cell labels. We observed that

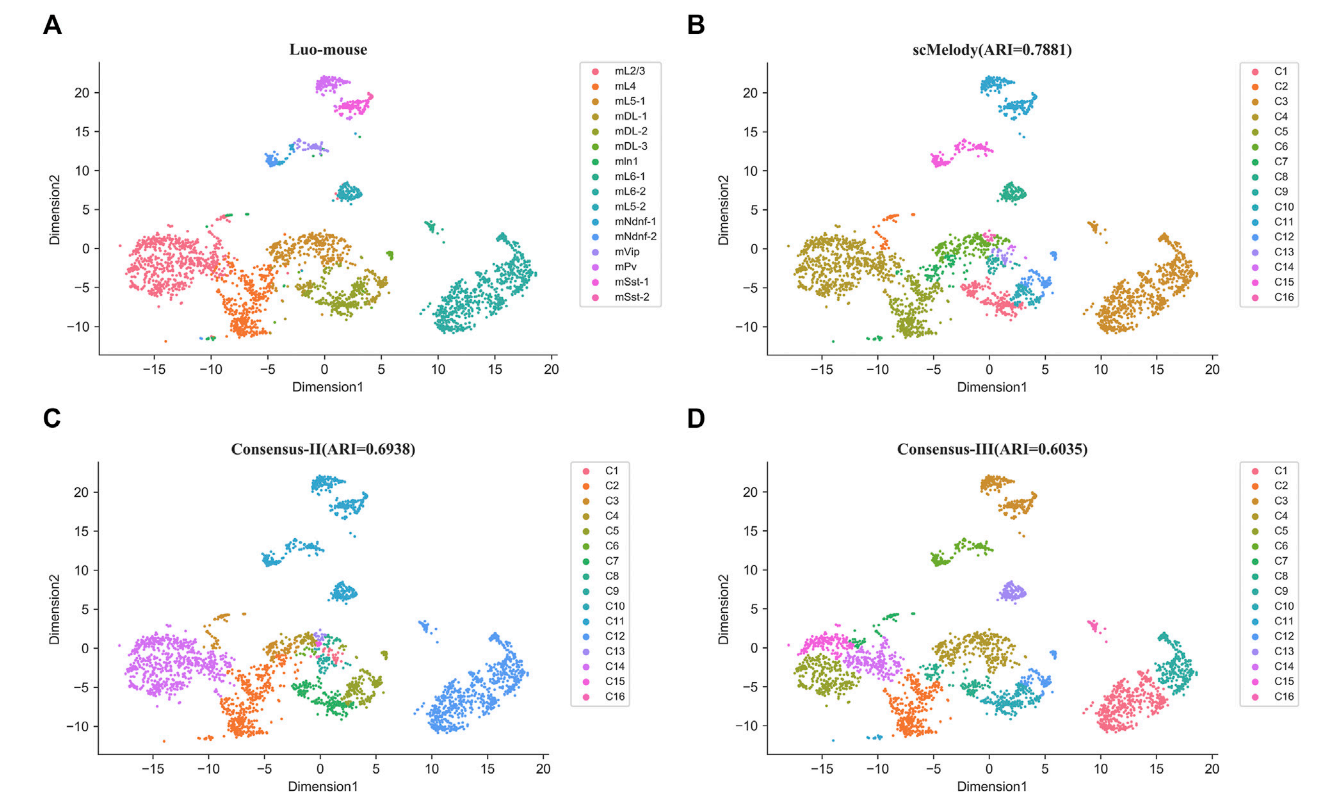


FIGURE 6 | t-SNE visualization results of the Luo-mouse dataset according to the different cell labels. **(A)** True cell labels; **(B)** The inferred clusters of scMeldoy; **(C)** The inferred clusters of Consensus-II. **(D)** The inferred clusters of Consensus-III.

the HSC population was more heterogeneous and a few HSCs presented a closer distance to MPPs. Moreover, **Figure 7C** showed the average methylation levels of the three groups of cells in the 500 most variable regions (Chi-square, FDR <.05) for each type of the regulatory region. According to Tukey's multiple comparisons test (Dunn, 1961), the average methylation level of the HSC-sub population was significantly different from that of the HSC-raw population in all six functional regions while was significantly different from that of the MPP population in four of six functional regions. The specific statistic information of the average methylation levels of the three groups of cells could be obtained in **Supplementary Tables S4–S9**. Moreover, we utilized the GREAT tool (McLean et al., 2010) to evaluate the functional significance of the identified variable genomic regions and the result indicated several enriched biological process (BP) Gene Ontology (GO) terms that were associated with HSC-raw and HSC-sub (**Figure 7D**; **Supplementary Table S10**). For instance, the two GO terms mitotic cytokinesis and positive regulation of mitotic nuclear division that were associated with hypomethylation in HSC-raw demonstrated that HSC-raw might have stronger differentiation potency than HSC-sub as DNA methylation could be associated with transcriptional repression (Luo et al., 2018). Finally, combined with the human hematopoietic lineage (Doulatov et al., 2012; Farlik et al., 2016), we knew that all blood cells originated from HSCs and the transition from HSC to MPP was always in the first stage of the differentiation lineage. These findings suggested that the six cells, which were annotated as HSC from the original publication, were different from the typical HSCs and presented an intermediate methylation status of two kinds of continuously differentiated cells (HSC and MPP) that warranted further investigation.

As an additional validation, we also evaluated the ability of scMelody to identify the novel cell clusters on a large dataset with complex cell composition contexts. This dataset was generated by Liu et al. (2021), in which there were 28077 inhibitory neurons derived from different regions of the mouse brain tissue, presenting high intercellular heterogeneity. We first aggregated the methylation profiles of 100 kb bins and these cells could be divided into 14 major types according to the annotations of the original experiment (**Figure 8A**). Besides, each major type was comprised of multiple heterogeneous subtypes, which were identified in the original experiment. When applying scMelody to this dataset, the clustering results showed that one major type PAL-Inh (inhibitory neurons derived from mouse pallidum) with the largest number of cells (4307 cells) among the 14 major types could be further divided into 11 subtypes, while only 10 subtypes were annotated for the PAL-Inh cells in the original experiment (**Figures 8B,C**). After comparison, we found that the novel subpopulation (PAL-Inh novel) identified by

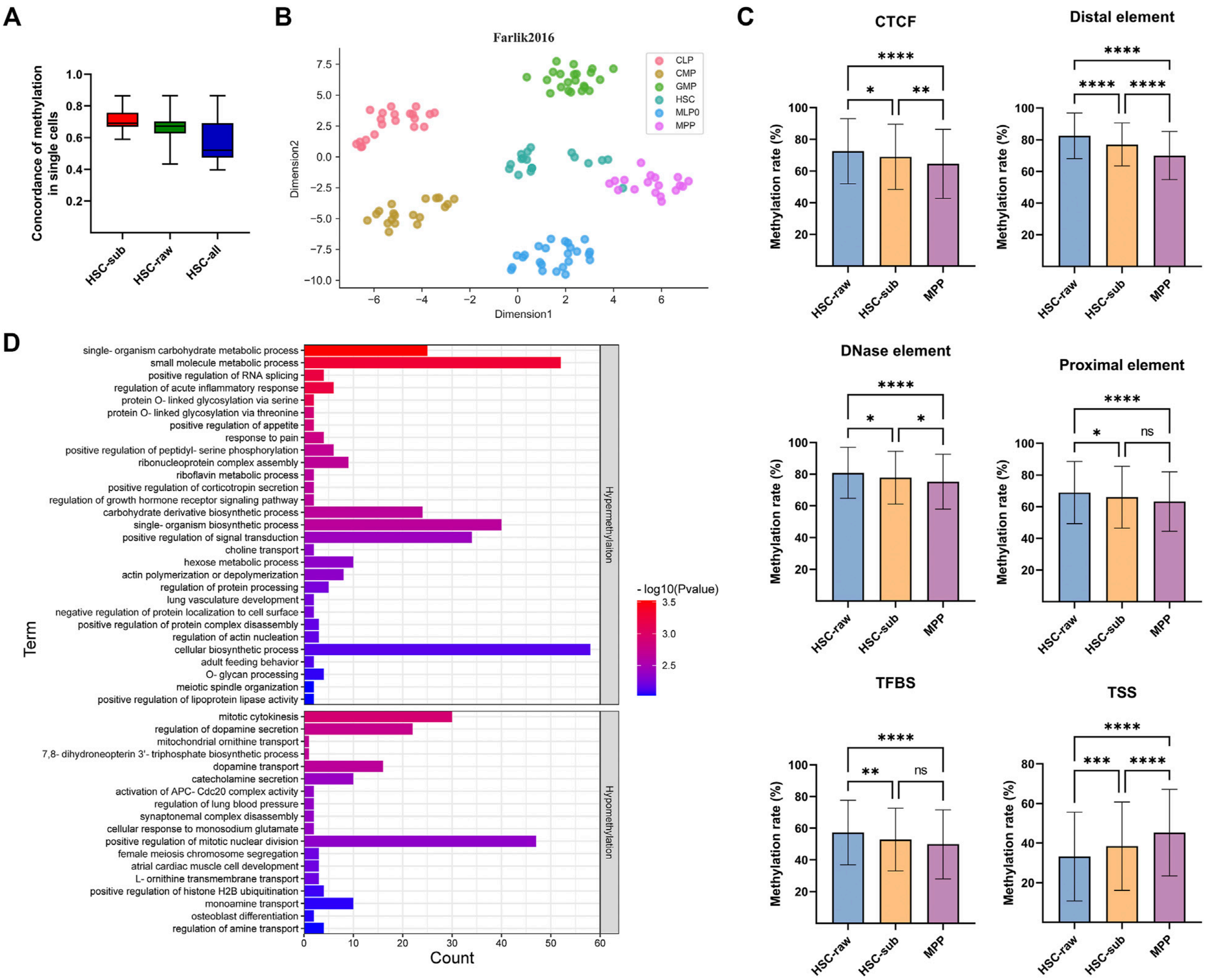


FIGURE 7 | Case study of scMelody in identifying novel cell clusters on the Farlik2016 dataset. **(A)** The concordance of the DNA methylation of the cells annotated as HSC. The concordance is calculated by averaging the pairwise correlation coefficients between any two single cells within each group, including Cosine, Hamming and Pearson correlation coefficient. **(B)** t-SNE projection plot of the Farlik2016 dataset using the average methylation levels on the top 500 variable functional regions in all six types of putative regulatory regions. Each point represents an individual cell, which is colored according to the annotated cell labels from the original experiment. **(C)** Average methylation levels of cells denoted as HSC-raw, HSC-sub and MPP on the six functional genomic regions, including CTCF binding site (CTCF), Distal element, DNase element, Proximal element, Transcription factor binding site (TFBS) and Transcriptional start site (TSS). Tukey's multiple comparisons test is used to determine whether there is a significant difference in mean methylation levels between each pair of the three cell groups. By default, the significance level is .05 and the significance marks are denoted by: ns, not significant; *p < .05; **p < .01, ***p < .001; ****p < .0001. **(D)** Genomic Regions Enrichment of Annotations Tool (GREAT) enrichment analysis of the variable genomic regions based on biological process Gene Ontology (GO) terms between the HSC-sub and HSC-raw. The enriched GO terms are ordered with the binomial test p value.

scMelody mainly came from the subtype PAL-Inh Meis2. Since the methylation levels on gene bodies negatively correlated with the gene expression in mouse neurons (Lister et al., 2013; Mo et al., 2015; Stroud et al., 2017; Liu et al., 2021), we profiled the methylation levels along the gene bodies with Chi-square (FDR <0.05) and the GO analysis revealed enriched BP terms for the differentially methylated genes between the PAL-Inh novel subpopulation and PAL-Inh Meis2 subpopulation (**Supplementary Figure S7**; **Supplementary Table S11**). For instance, several most significantly enriched GO terms, such as nervous system development and neurogenesis, clearly showed major biological processes of mouse neuron development. Moreover, we also noticed that the GO term "cell morphogenesis involved in neuron differentiation" was associated with hypermethylation in PAL-Inh novel subpopulation and the GO term "negative regulation of protein modification process" was associated with hypomethylation in PAL-Inh novel subpopulation. This result showed that the PAL-Inh Meis2 subpopulation might have a stronger differentiation ability than the PAL-Inh novel subpopulation (Menon and Gupton, 2018; Badimon et al., 2020). Besides, the GREAT analysis uncovered the term "abnormal neuron morphology" of

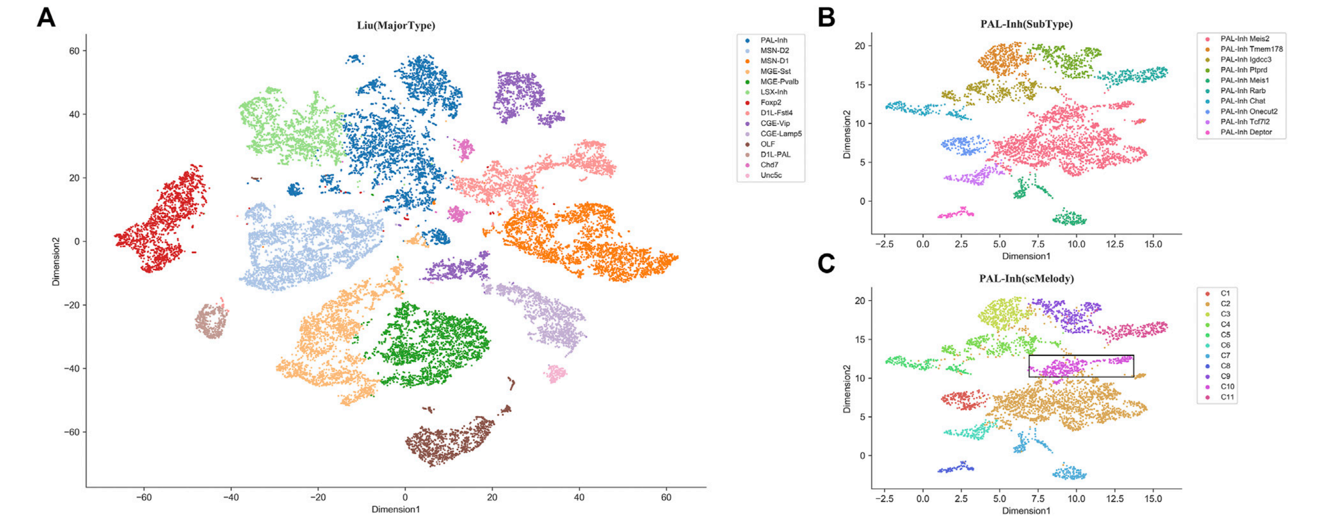


FIGURE 8 | t-SNE visualization results of the large Liu dataset based on the 100 kb bins methylation profiles. **(A)** The t-SNE visualization result of all inhibitory neurons, where a total of 28077 cells are defined as 14 major types and are colored according to the annotations from the original experiment. **(B)** The t-SNE visualization result of PAL-Inh subpopulation, where a total of 4307 cells are defined as 10 subtypes are colored according to the annotations from the original experiment. **(C)** The t-SNE visualization result of PAL-Inh subpopulation, where the cells are clustered into 11 subtypes by scMelody. For comparison, the novel cell cluster identified by scMelody is circled with a black rectangle.

Mouse Phenotype, which further confirmed the difference in these two cell subpopulations.

4 DISCUSSION

The high resolution of single-cell methylation sequencing enables researchers to explore cell-to-cell epigenetic heterogeneity and underlines the significance of clustering cells based on the single-cell methylation profiles. In a biological sense, DNA methylation is well suited for exploring cell heterogeneity because this crucial modification is cell-type-specific and preserves an epigenetic memory of a cell's developmental history (Farlik et al., 2016). In this paper, we propose scMelody, an enhanced consensus-based clustering model for single-cell methylation data analysis by reconstructing cell-to-cell pairwise similarity. When applying it on real single-cell methylation datasets generated from various sequencing techniques, scMeldoy achieved significant clustering performance gains over the previous methods, including several single-distance-based methods and one probabilistic method. Benefiting from the reconstructed cell-to-cell similarity measure, scMelody also attained accurate estimates for the number of clusters based on the silhouette coefficient criterion. Moreover, using the synthetic datasets generated across a variety of settings, scMelody was demonstrated to be stable which robustly maintained its clustering accuracy over a wide range of number of cells, number of clusters and CpG dropout proportions. The real case studies also indicated the capability of scMelody to identify known cell types and uncover novel cell clusters. To sum up, scMeldoy could accurately recapitulate the cellular epigenetic heterogeneity and was demonstrated to be universal for different kinds of single-cell methylation datasets.

Generally, the (dis)similarity measure is the core for quantifying the methylation differences between cells, thus many methods are designed to incorporate different cell-to-cell methylation (dis) similarity measures into the distance-based clustering algorithms to generate cell partitions. However, our results showed that no single (dis)similarity measure could provide satisfactory clustering performance on all datasets as different (dis)similarity measures captures the cellular heterogeneity from different perspectives. For example, both PearsonHC and PDclust accurately assigned all cells to their respective clusters on the Pott dataset while they could hardly identify the cell types on the Farlik2016 dataset (**Figure 4**). Instead, a significant advantage of scMelody was that it integrated the clustering information of multiple basic similarity measures to overcome their limitation in capturing complete cellular methylation heterogeneity. Besides, the reconstructed cell-to-cell similarity measure enabled scMelody to reach better clustering performance across different datasets. This highlighted the importance of identifying cell subpopulations by combining the information of different cell-to-cell methylation (dis)similarity relationships. However, even scMelody can process thousands of cells within several hours, the computational efficiency of scMelody is still to be improved especially when the computational resources are limited. We will continue to develop optimized versions of scMelody to improve its computational efficiency, such as the GPU-accelerated scMelody, which can be more practical for the researchers to use it.

With the development of single-cell methylation sequencing technologies, the increase of sequencing depth will greatly alleviate the sparsity problem of single-cell methylation data, which can significantly boost the performance of clustering cells based on cell-to-cell similarity patterns. Our scMelody is flexible and can easily accommodate additional similarity measures to cluster cells, as the novel and sophisticated distance measures continue to be proposed. This has important implications for fully utilizing single-cell methylation sequencing to study cell differentiation versus variation, especially for uncovering novel cell types in complex human diseases, such as cancers.

DATA AVAILABILITY STATEMENT

An implementation of scMelody is freely available at https://github.com/TQBio/scMelody. All datasets used in this paper can be obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The synthetic datasets are generated from the bulk RRBS data with GEO accession number GSE27584 and the real single-cell methylation datasets analyzed in this paper can be obtained with the corresponding GEO accession numbers (**Table 1**). The large Liu dataset can be obtained with GEO accession number GSE132489.

AUTHOR CONTRIBUTIONS

QT designed the method carried out in the study. JT collected the real datasets and generated the synthetic datasets. QT, JZ, LL, XC, and SF performed the analysis. QT edited the manuscript. JZ and SF led the research and reviewed the manuscript. All authors read and approved the manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China (no. 61872063), Shenzhen Science and Technology Program, China (no. JCYJ20210324140407021) and Grant SCITLAB-0006 of intelligent Terminal Key Laboratory of Sichuan Province, China.

ACKNOWLEDGMENTS

We would like to thank the associate editor and the reviewers for their valuable comments and help that improved this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.842019/full#supplementary-material

REFERENCES

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., et al. (2012). BLUEPRINT to Decode the Epigenetic Signature Written in Blood. *Nat. Biotechnol.* 30, 224–226. doi:10.1038/nbt.2153

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., et al. (2016). Parallel Single-Cell Sequencing Links Transcriptional and Epigenetic Heterogeneity. *Nat. Methods* 13, 229–232. doi:10.1038/nmeth.3728

Aran, D., and Hellman, A. (2013). DNA Methylation of Transcriptional Enhancers and Cancer Predisposition. *Cell* 154, 11–13. doi:10.1016/j.cell.2013.06.018

Badimon, A., Strasburger, H. J., Ayata, P., Chen, X., Nair, A., Ikegami, A., et al. (2020). Negative Feedback Control of Neuronal Activity by Microglia. *Nature* 586, 417–423. doi:10.1038/s41586-020-2777-8

Boongoen, T., and Iam-On, N. (2018). Cluster Ensembles: A Survey of Approaches with Recent Extensions and Applications. *Comput. Sci. Rev.* 28, 1–25. doi:10.1016/j.cosrev.2018.01.003

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., et al. (2008). Shotgun Bisulphite Sequencing of the Arabidopsis Genome Reveals DNA Methylation Patterning. *Nature* 452, 215–219. doi:10.1038/nature06745

Cui, Y., Zhang, S., Liang, Y., Wang, X., Ferraro, T. N., and Chen, Y. J. B. I. B. (2021). Consensus Clustering of Single-Cell RNA-Seq Data by Enhancing Network Affinity. *Brief. Bioinform.* 22, bbab236. doi:10.1093/bib/bbab236

de Souza, C. P. E, Andronescu, M., Masud, T., Kabeer, F., Biele, J., Laks, E., et al. (2020). Epiclomal: Probabilistic Clustering of Sparse Single-Cell DNA Methylation Data. *Plos Comput. Biol.* 16, e1008270. doi:10.1371/journal.pcbi.1008270

Doulatov, S., Notta, F., Laurenti, E., and Dick, J. E. (2012). Hematopoiesis: A Human Perspective. *Cell Stem Cell* 10, 120–136. doi:10.1016/j.stem.2012.01.006

Dunn, O. J. (1961). Multiple Comparisons Among Means. *J. Am. Stat. Assoc.* 56, 52–64. doi:10.1080/01621459.1961.10482090

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., et al. (2015). Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cel Rep.* 10, 1386–1397. doi:10.1016/j.celrep.2015.02.001

Farlik, M., Halbritter, F., Müller, F., Choudry, F. A., Ebert, P., Klughammer, J., et al. (2016). DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell* 19, 808–822. doi:10.1016/j.stem.2016.10.019

Fern, X. Z., Lin, W., and Journal, D. M. T. A. D. S. (2008). Cluster Ensemble Selection. *Stat. Analy Data Mining* 1, 128–141. doi:10.1002/sam.10008

Ghaemi, R., Sulaiman, M. N., Ibrahim, H., and Mustapha, N. J. W. A. O. S. (2009). Engineering, and Technology. *A Surv. Clustering Ensembles Tech.* 50, 636–645. doi:10.5281/zenodo.1329276

Golalipour, K., Akbari, E., Hamidi, S. S., Lee, M., and Enayatifar, R. (2021). From Clustering to Clustering Ensemble Selection: A Review. *Eng. Appl. Artif. Intel* 104, 104388. doi:10.1016/j.engappai.2021.104388

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-Cell Methylome Landscapes of Mouse Embryonic Stem Cells and Early Embryos Analyzed Using Reduced Representation Bisulfite Sequencing. *Genome Res.* 23, 2126–2135. doi:10.1101/gr.161679.113

Hadjitodorov, S. T., Kuncheva, L. I., and Todorova, L. P. (2006). Moderate Diversity for Better Cluster Ensembles. *Inf. Fusion* 7, 264–275. doi:10.1016/j.inffus.2005.01.008

Haranczyk, M., and Holliday, J. (2008). Comparison of Similarity Coefficients for Clustering and Compound Selection. *J. Chem. Inf. Model.* 48, 498–508. doi:10.1021/ci700413a

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., et al. (2016). Single-Cell Triple Omics Sequencing Reveals Genetic, Epigenetic, and Transcriptomic Heterogeneity in Hepatocellular Carcinomas. *Cell Res* 26, 304–319. doi:10.1038/cr.2016.23

Hubert, L., and Arabie, P. (1985). Comparing Partitions. *J. Classification* 2, 193–218. doi:10.1007/bf01908075

Hui, T., Cao, Q., Wegrzyn-Woltosz, J., O'Neill, K., Hammond, C. A., Knapp, D. J. H. F., et al. (2018). High-Resolution Single-Cell DNA Methylation Measurements Reveal Epigenetically Distinct Hematopoietic Stem Cell Subpopulations. *Stem Cel Rep.* 11, 578–592. doi:10.1016/j.stemcr.2018.07.003

Kapourani, C.-A., and Sanguinetti, G. (2019). Melissa: Bayesian Clustering and Imputation of Single-Cell Methylomes. *Genome Biol.* 20, 61. doi:10.1186/s13059-019-1665-8

Kapourani, C. A., Argelaguet, R., Sanguinetti, G., and Vallejos, C. A. (2021). scMET: Bayesian Modeling of DNA Methylation Heterogeneity at Single-Cell Resolution. *Genome Biol.* 22, 114. doi:10.1186/s13059-021-02329-8

Khalifa, A. A., Haranczyk, M., and Holliday, J. (2009). Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection. *J. Chem. Inf. Model.* 49, 1193–1201. doi:10.1021/ci8004644

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: Consensus Clustering of Single-Cell RNA-Seq Data. *Nat. Methods* 14, 483–486. doi:10.1038/nmeth.4236

Koch, A., Joosten, S. C., Feng, Z., de Ruijter, T. C., Draht, M. X., Melotte, V., et al. (2018). Analysis of DNA Methylation in Cancer: Location Revisited. *Nat. Rev. Clin. Oncol.* 15, 459–466. doi:10.1038/s41571-018-0004-4

Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). DNA Methylome Analysis Using Short Bisulfite Sequencing Data. *Nat. Methods* 9, 145–151. doi:10.1038/nmeth.1828

Kuncheva, L. I., and Hadjitodorov, S. T. (2004). "Using Diversity in Cluster Ensembles," in Ieee Sys Man Cybern, The Hague, Netheirlands, 10–13 October, 2004, 1214–1219.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., et al. (2013). Global Epigenomic Reconfiguration during Mammalian Brain Development. *Science* 341, 1237905. doi:10.1126/science.1237905

Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., et al. (2021). DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution. *Nature* 598, 120–128. doi:10.1038/s41586-020-03182-8

Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., et al. (2017). Single-Cell Methylomes Identify Neuronal Subtypes and Regulatory Elements in Mammalian Cortex. *Science* 357, 600–604. doi:10.1126/science.aan3351

Luo, C., Hajkova, P., and Ecker, J. R. (2018). Dynamic DNA Methylation: In the Right Place at the Right Time. *Science* 361, 1336–1340. doi:10.1126/science.aat6806

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., et al. (2010). GREAT Improves Functional Interpretation of Cis-Regulatory Regions. *Nat. Biotechnol.* 28, 495–501. doi:10.1038/nbt.1630

Menon, S., and Gupton, S. J. F. (2018). Recent Advances in Branching Mechanisms Underlying Neuronal Morphogenesis. *F1000Res* 7, F1000. doi:10.12688/f1000research.16038.1

Mo, A., Mukamel, E. A., Davis, F. P., Luo, C., Henry, G. L., Picard, S., et al. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron* 86, 1369–1384. doi:10.1016/j.neuron.2015.05.018

Oakes, C. C., Seifert, M., Assenov, Y., Gu, L., Przekopowitz, M., Ruppert, A. S., et al. (2016). DNA Methylation Dynamics during B Cell Maturation Underlie a Continuum of Disease Phenotypes in Chronic Lymphocytic Leukemia. *Nat. Genet.* 48, 253–264. doi:10.1038/ng.3488

Pott, S. (2017). Simultaneous Measurement of Chromatin Accessibility, DNA Methylation, and Nucleosome Phasing in Single Cells. *Elife* 6, e23203. doi:10.7554/eLife.23203

Rosenberg, A., and Hirschberg, J. (2007). "V-measure: A Conditional Entropy-Based External Cluster Evaluation Measure," in Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June, 2007, 410–420.

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Maths.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., et al. (2011). Validation of a DNA Methylation Microarray for 450,000 CpG Sites in the Human Genome. *Epigenetics* 6, 692–702. doi:10.4161/epi.6.6.16196

Schwartzman, O., and Tanay, A. (2015). Single-Cell Epigenomics: Techniques and Emerging Applications. *Nat. Rev. Genet.* 16, 716–726. doi:10.1038/nrg3980

Shirkhorshidi, A. S., Aghabozorgi, S., and Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *Plos One* 10, e0144059. doi:10.1371/journal.pone.0144059

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

Strehl, A., and Ghosh, J. (2002). Cluster Ensembles---a knowledge reuse framework for combining multiple partitions. *J. Machine Learn. Res.* 3, 583–617. doi:10.1162/153244303321897735

Stroud, H., Su, S. C., Hrvatin, S., Greben, A. W., Renthal, W., Boxer, L. D., et al. (2017). Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. *Cell* 171, 1151–1164. doi:10.1016/j.cell.2017.09.047

Ünlü, R., and Xanthopoulos, P. (2019). A Weighted framework for Unsupervised Ensemble Learning Based on Internal Quality Measures. *Ann. Oper. Res.* 276, 229–247. doi:10.1007/s10479-017-2716-8

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach Learn. Res.* 9, 2579–2605.

Vega-Pons, S., and Ruiz-Shulcloper, J. (2011). A Survey of Clustering Ensemble Algorithms. *Int. J. Patt. Recogn. Artif. Intell.* 25, 337–372. doi:10.1142/s0218001411008683

Vega-Pons, S., Correa-Morris, J., and Ruiz-Shulcloper, J. (2008). Weighted Cluster Ensemble Using a Kernel Consensus FunctionProgress in Pattern Recognition, Image Analysis and Applications. *Proceedings* 5197, 195–202. doi:10.1007/978-3-540-85920-8_24

Vega-Pons, S., Ruiz-Shulcloper, J., and Guerra-Gandón, A. (2011). Weighted association based methods for the combination of heterogeneous partitions. *Pattern Recognition Lett.* 32, 2163–2170. doi:10.1016/j.patrec.2011.05.006

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach Learn. Res.* 11, 2837–2854. doi:10.1145/1553374.1553511

von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *Stat. Comput.* 17, 395–416. doi:10.1007/s11222-007-9033-z

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., et al. (2012). Widespread Plasticity in CTCF Occupancy Linked to DNA Methylation. *Genome Res.* 22, 1680–1688. doi:10.1101/gr.136101.111

Wang, C., Mu, Z., Mou, C., Zheng, H., and Liu, J. (2021). Consensus-Based Clustering of Single Cells by Reconstructing Cell-to-Cell Dissimilarity. *Brief. Bioinform.* 23, bbab379. doi:10.1093/bib/bbab379

Yona, G., Dirks, W., Rahman, S., and Lin, D. M. (2006). Effective Similarity Measures for Expression Profiles. *Bioinformatics* 22, 1616–1622. doi:10.1093/bioinformatics/btl127

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., and Flicek, P. R. (2015). The Ensembl Regulatory Build. *Genome Biol.* 16, 56. doi:10.1186/s13059-015-0621-5

Zhang, M. (2021). Weighted Clustering Ensemble: A Review. *Pattern Recognition*, 108428. doi:10.1016/j.patcog.2021.108428

Zhu, X. S., Li, J., Li, H. D., Xie, M., and Wang, J. X. (2020). Sc-GPE: A Graph Partitioning-Based Cluster Ensemble Method for Single-Cell. *Front. Genet.* 11, 604790. doi:10.3389/fgene.2020.604790

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.

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fbioe.2022.883791

Detection and Localization of Solid Tumors Utilizing the Cancer-Type-Specific Mutational Signatures

Ziyu Wang[1,2,3†], *Tingting Zhang*[1,2,3†], *Wei Wu*[1,2,3†], *Lingxiang Wu*[1,2,3†], *Jie Li*[1,2,3], *Bin Huang*[1,2,3], *Yuan Liang*[1,2,3], *Yan Li*[1,2,3], *Pengping Li*[1,2,3], *Kening Li*[1,2,3]*, *Wei Wang*[4]*, *Renhua Guo*[5]* *and Qianghu Wang*[1,2,3]*

[1]Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research, The Affiliated Cancer Hospital of Nanjing Medical University, Nanjing, China, [2]Department of Bioinformatics, Nanjing Medical University, Nanjing, China, [3]Institute for Brain Tumors, Jiangsu Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China, [4]Department of Thoracic Surgery, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China, [5]Department of Oncology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:
Zhaoyuan Fang,
Zhejiang University, China

Reviewed by:
Lin Zhang,
University of California, San Diego, United States
Siyuan Zheng,
The University of Texas Health Science Center at San Antonio, United States

***Correspondence:**
Kening Li
likening@njmu.edu.cn
Wei Wang
wangwei15261883958@163.com
Renhua Guo
rhguo@njmu.edu.cn
Qianghu Wang
wangqh@njmu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *25 February 2022*
Accepted: *07 April 2022*
Published: *25 April 2022*

Citation:
Wang Z, Zhang T, Wu W, Wu L, Li J, Huang B, Liang Y, Li Y, Li P, Li K, Wang W, Guo R and Wang Q (2022) Detection and Localization of Solid Tumors Utilizing the Cancer-Type-Specific Mutational Signatures. Front. Bioeng. Biotechnol. 10:883791. doi: 10.3389/fbioe.2022.883791

Accurate detection and location of tumor lesions are essential for improving the diagnosis and personalized cancer therapy. However, the diagnosis of lesions with fuzzy histology is mainly dependent on experiences and with low accuracy and efficiency. Here, we developed a logistic regression model based on mutational signatures (MS) for each cancer type to trace the tumor origin. We observed MS could distinguish cancer from inflammation and healthy individuals. By collecting extensive datasets of samples from ten tumor types in the training cohort (5,001 samples) and independent testing cohort (2,580 samples), cancer-type-specific MS patterns (CTS-MS) were identified and had a robust performance in distinguishing different types of primary and metastatic solid tumors (AUC: 0.76 ~ 0.93). Moreover, we validated our model in an Asian population and found that the AUC of our model in predicting the tumor origin of the Asian population was higher than 0.7. The metastatic tumor lesions inherited the MS pattern of the primary tumor, suggesting the capability of MS in identifying the tissue-of-origin for metastatic cancers. Furthermore, we distinguished breast cancer and prostate cancer with 90% accuracy by combining somatic mutations and CTS-MS from cfDNA, indicating that the CTS-MS could improve the accuracy of cancer-type prediction by cfDNA. In summary, our study demonstrated that MS was a novel reliable biomarker for diagnosing solid tumors and provided new insights into predicting tissue-of-origin.

Keywords: cancer biomarkers, cancer diagnosis, cancer localization, mutational signatures, liquid biopsy

INTRODUCTION

An accurate cancer diagnosis is crucial for choosing the optimal therapy and predicting clinical outcomes (Jerjes et al., 2010; Varadhachary and Raber, 2014; Thomson, 2018). Histological examination of the resected specimen remains the gold standard for diagnosing tumors. However, rapid, accurate diagnosis based on morphology and routine ancillary techniques is challenging for lesions with fuzzy histology, especially metastatic cancers (Saudemont et al., 2018; Conway et al., 2019). The accuracies of computed tomography and positron emission tomography in identifying the tissue-of-origin of the carcinoma with unknown primary were 20–27% and 24–40%, respectively, which are far from enough for determining targeted therapies (Fu et al., 2019; He et al., 2020a). Therefore, effective strategies are urgently needed for tumor detection and localization.

The mutation data is easily accessible molecular profile, which could be robustly retrieved and sequenced in various samples, such as formalin-fixed and paraffin-embedded specimens. Previous studies showed a high concordance in mutational patterns between primary and metastatic tumors, especially when pathogenic mutations in driver genes were considered (Manca et al., 2019). Accordingly, some methods were proposed for tumor origin prediction based on somatic mutations (Dietlein and Eschner, 2014; Marquard et al., 2015; Jiao et al., 2020). However, somatic mutations also could be detected in healthy individuals (Welch et al., 2012; Blokzijl et al., 2016; Martincorena and Campbell, 2016), increasing the difficulty of cancer diagnosis. Moreover, mutational profiles showed substantial overlap across different cancer types, making it difficult to trace the origin of the tumor (Jurmeister et al., 2019).

Somatic mutations result from multiple mutational processes, including exposure to exogenous or endogenous mutagens, enzymatic modification of DNA, and defective DNA repair. Different mutational processes generate unique combinations of mutation types, termed mutational signatures (MS). Single nucleotide variants can be divided into six types according to the type of base substitution: C > A, C > G, C > T, T > A, T > C, T > G. Alexandrov et al. extended the original classification of six types of single-base substitutions by including base 5′ and base 3′ to the somatic mutation. Mutational signature (MS) is created by counting the number of substitutions for each of these 96 mutation types. The COSMIC database has described 30 types of reference MS based on the analyses of ~10,000 whole-genome or whole-exon sequencing datasets from TCGA and ICGC databases (https://cancer.sanger.ac.uk/signatures/signatures_v2/). MS is cancer-derived etiologies that provide a powerful alternative for understanding cancer pathophysiology (Alexandrov et al., 2013; Helleday et al., 2014; Roberts and Gordenin, 2014; Alexandrov et al., 2016; Pilati et al., 2017; Zou et al., 2018). Unlike the extensive heterogeneity of somatic mutations across samples, MS is more stable across individuals in the same tumor type. Previous studies reported that different tumor types leave distinctive patterns of MS (Degasperi et al., 2020). For example, the MS patterns generated in experimental systems for tobacco carcinogens exposure were observed in lung cancer (Alexandrov et al., 2016). MS patterns in colorectal cancer are mostly related to defective DNA mismatch repair (Pandey et al., 2019). Therefore, we reasonably speculated that MS patterns could predict the tumor origin.

Based on the MS patterns, we used the logistic regression method to construct a model for each cancer type to predict the origin. Our results showed that MS could distinguish cancer patients from healthy individuals and inflammation. Furthermore, our MS-based models showed high accuracy in detecting the origin of tumors in both primary and metastatic lesions. Besides, we also found that MS had a better performance in distinguishing various cancer types than somatic mutations. Finally, we indicated that considering the MS patterns could help increase the accuracy of cancer-type prediction by cfDNA.

MATERIALS AND METHODS

Collection of the Whole Exome Data of Tissues and cfDNA

All variant data of primary tumors were downloaded from TCGA (http://gdac.broadinstitute.org/), International cancer genome consortium (ICGC, https://icgc.org/), and other previous studies (**Supplementary Tables S1, S2**). In these cases, we used only the data in TCGA for training (Data Set1). The data outside of TCGA were validated (Data set 2). The somatic profiles of metastatic tumors were derived from 303 metastatic tumors across nine tumor types (**Supplementary Table S3**). We assembled several sets of normal or inflammatory tissues to evaluate the difference in genomic landscape between tumor patients and healthy individuals. One of the data sets included 28 healthy individuals, 48 patients with ulcerative colitis, and 18 patients with colitis-associated neoplasia, and the other data set contained 9 normal brains tissues, 13 normal colon tissues, and 13 normal kidney tissues. We also acquired somatic mutations from 27 breast and 14 prostate cancers of cfDNA and biopsy. All these data were obtained by whole-exome sequencing and aligned to the hg19 genome.

Identification of the Cancer-Type-Specific Mutational Signatures Patterns

The characteristic MS patterns of each cancer type meet the following requirements. First, MS was observed in at least 20% of samples. Secondly, there were significant differences compared with other cancer types, including a fold change greater than 1.5 and an absolute difference greater than 0.1.

Mutational Signatures-Based Machine Learning Procedure for Predicting the Cancer Types of the Primary Tumor

For each of the ten cancer types selected from the TCGA data set, we used a stepwise logistic regression model to train classifiers for each cancer type on the CTS-MS described in the above section and validated our models in an independent dataset. To evaluate the performance of our model in different populations, we downloaded the somatic mutation data for Asian populations from the ICGC database, including non-small cell lung cancers (n = 76), colorectal cancers (n = 187), bladder cancers (n = 103), gastric cancers (n = 10), and liver cancers (n = 163). We developed a logistic regression model based on MS for each cancer type to trace the tumor origin. Take breast cancer as an instance, we calculated the score of each sample in the validation dataset using the breast cancer model, labeling breast cancer patients as “1” and non-breast cancer patients as “0” to obtain grouping information. The prediction performance of AUC was calculated using the predicted values estimated by the model with the combination of selected MS as predictors and the group as an outcome.

Tracing the Origin of Metastatic Sites Based on Mutational Signatures Patterns

First, we used the liver cancer model above to distinguish primary liver tumors and malignant liver lesions originating from other tissues. We further predicted the origin of lesions originating from other tissues, which were correctly classified in the previous step, including 28 breast cancers, 9 esophagus cancers, and 10 prostate cancers. To predict the origin of malignant liver lesions originating from other tissues, we combined CTS-MS and the score of these three primary tumor models to train a classifier by neural networks based on the three cancer types selected from the TCGA data set. Then, we used the model to predict the origin of malignant liver lesions originating from other tissues.

Combination of Mutational Signatures Patterns and Somatic Mutation to Distinguish Different Cancer Types Based on Plasma cfDNA Data

Based on the CTS-MS, we predicted the origin of tumors from cfDNA, including 27 breast cancers and 14 prostate cancers. We compared the scores of each sample in the breast cancer model and the prostate cancer model. The origin of the sample was considered to be from the tumor type with a high score. Further, we combined CTS-MS and tumor-specific mutations to improve the precision. We identified the tumor-specific mutations as follows: 1) we calculated the frequency of mutations in each gene in each cancer and identified genes that were mutated in more than 5% of the samples as candidate markers. 2) it was considered a tumor-specific mutation if the mutation frequency changes more than 0.1 compared with other cancer types. Then, using the stepwise logistic regression model, we developed classifiers for prostate and breast cancer based on the CTS-MS and tumor-specific mutations.

Statistical Analysis

The deconstructSigs approach was used to determine the linear combination of pre-defined signatures of a single tumor sample (Rosenthal et al., 2016). We next applied SomaticSignatrues to identify the *de novo* MS (Gehring et al., 2015). The information of pre-defined MS was downloaded from the COSMIC database. The *de novo* MS was mapped to pre-defined MS through cosine similarity. If the similarity was higher than 0.75, it was considered the same MS.

We annotated the mutated genes in each sample in the STRING database (https://string-db.org/). According to the STRING database, we constructed a network of protein-protein interactions for all mutated genes in each sample. Mutation connection scores were defined by gene connectivity, measured by the ratio of the number of genes with interactions to the total number of mutated genes (**Eq. 1**). Larger mutation connection scores indicate that the mutated gene is more functionally relevant.

$$Mutation\,connection\,score = \frac{the\,number\,of\ genes\,with\,interactions}{the\,total\,number\,of\ mutated\,genes} \tag{1}$$

We calculated the similarity between tumors as **Eq. 2**. For each sample i of tumor M and each sample j of tumor N, we calculated the cosine similarity (rho) between i and j based on pre-defined MS. Finally, a similarity matrix with m rows and n columns was generated. We performed zero-mean normalization on each row and each column of the similarity matrix. Then, we ranked each row and divided it by the number of columns. Further, we ranked each column and divided it by the number of rows.

$$similarity = \sum\nolimits_{i \le m} \frac{\sum\nolimits_{j \le n} rho(i, j)^2}{mn} \tag{2}$$

Statistical analyses were performed using R software. The significance probability (p) values were calculated by the two-tailed Wilcoxon test functions in R, and the LSAfun package calculated the cosine similarity. Figures were drawn using the ggplot2, or package under R environment.

RESULTS

Mutational Signatures Patterns Distinguish Cancers From Inflammation and Healthy Individuals

To compare the difference in the genomic landscape among tumor patients, non-tumor inflammation patients, and healthy individuals, we collected three datasets, including healthy individuals (HI, n = 28), patients with ulcerative colitis (UC, n = 48), and patients with colitis-associated neoplasia (CAN, n = 18) (Nanki et al., 2020). We first computationally defined a tumor mutation connection score measurement to explore whether the mutated genes were functionally related. The higher the tumor mutation connection score, the stronger the functional relevance of the mutated genes in the individual (detail in methods). Results showed that the functional relevance of the mutated genes in CAN is significantly different from HI and UC (**Figure 1A**). The tumor mutation connection score of CAN was significantly higher than HI and UC (HI vs. CAN, Wilcoxon rank-sum test p < 0.001; UC vs. CAN, Wilcoxon rank-sum test p < 0.001), indicating that rather than randomly mutation, specific endogenous or exogenous factors were involved in the mutation genesis in CAN. Accordingly, we next explored the potential causal factors of the differences between CAN and HI/UC. Using the non-negative matrix factorization method, we identified two known MS that showed differential contributions among cancer, normal, and inflammation groups (**Figures 1B,C; Supplementary Figure S1A**), one of which is related to aging and the other is associated with DNA mismatch repair defective (MMR). The contribution of aging-related MS was remarkably higher in CAN than in HI and UC (HI vs. CAN, Wilcoxon rank-

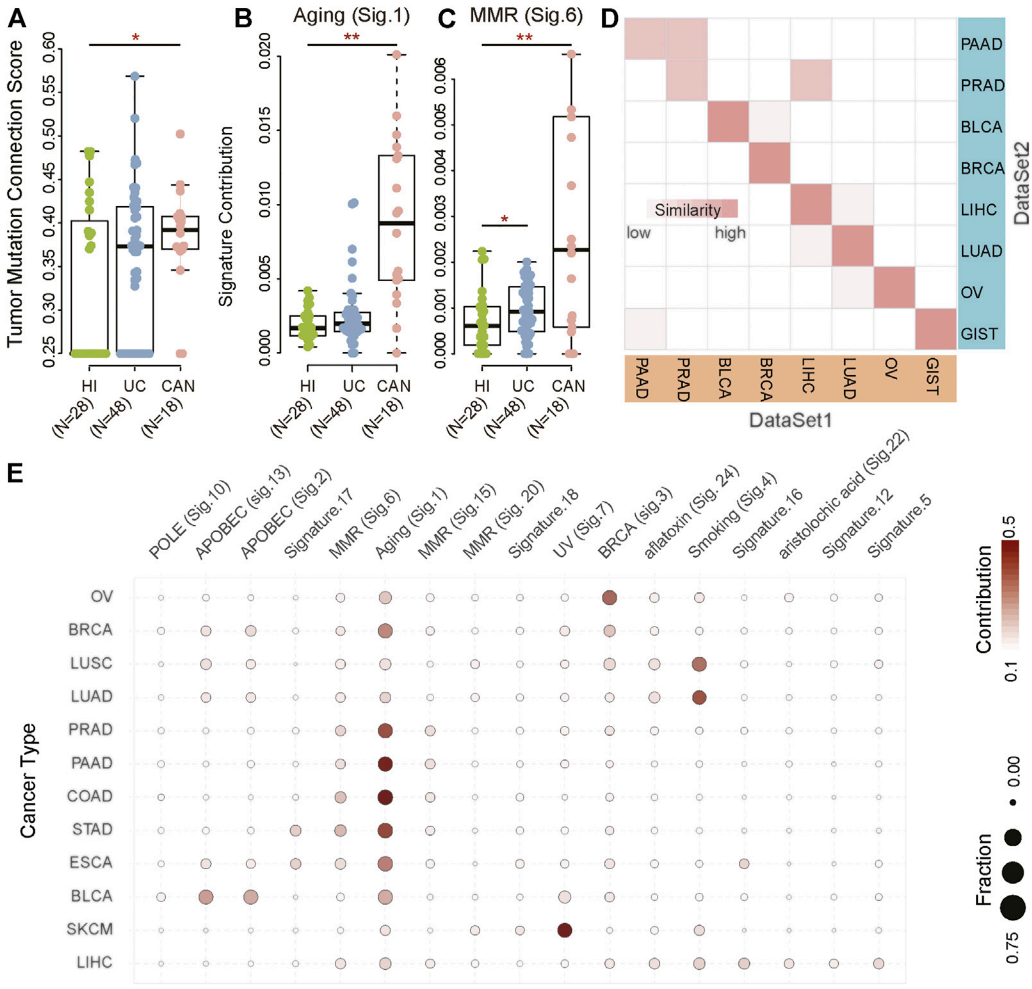


FIGURE 1 | Mutational signatures for cancer diagnosis. **(A–C)** The biological processes of accumulated mutations in healthy individuals (HI) and patients with ulcerative colitis (UC) and colitis-associated neoplasia (CAN). **(D)** The correlation between DataSet1 (TCGA) and DataSet2 (previous studies) based on MS in bladder cancer (BLCA), non-small cell lung cancer (NSCLC), pancreatic cancer (PAAD), breast cancer (BRCA), ovarian serous cystadenocarcinoma (OV), liver hepatocellular carcinoma (LIHC), and gastrointestinal cancer, including colorectal cancer (CRC), esophageal carcinoma (ESCA), and stomach adenocarcinoma (STAD). The darker the color, the higher the similarity. **(E)** Heatmaps of MS in BLCA ($n = 412$), NSCLC ($n = 1,108$), PAAD ($n = 179$), BRCA ($n = 985$), OV ($n = 435$), skin cutaneous melanoma (SKCM, $n = 468$), LIHC ($n = 464$), CRC ($n = 398$), ESCA ($n = 184$), and STAD ($n = 439$). The color indicates the average contribution of MS. The size of the dots indicates the fraction. Fraction: The proportion of samples with a mutational signature contribution of more than 0.06 in each cancer type as a proportion of the total samples. Contribution: Average contribution of each mutational signature in each cancer type.

sum test $p < 0.001$; UC vs. CAN, Wilcoxon rank-sum test $p < 0.001$, **Figure 1B**). To avoid bias from age, we checked the distribution of age across three groups in our dataset. There were no differences in the age distribution of the tumor and healthy individuals/inflammation patients (ANOVA test, $p = 0.319$, **Supplementary Figure S1B**). Furthermore, we found that senior individuals were biased towards higher age-related signature in the healthy population (Spearman correlation:0.41, $p = 0.035$, **Supplementary Figures S1C,D**). However, there was no correlation between age and the age-related signature of cancer patients (Spearman correlation -0.16, $p = 0.51$, **Supplementary Figures S1C,D**). Notably, the age-related signature of tumor patients was much higher than those of healthy individuals/inflammatory patients across all age groups. Even the weights of age-related signature in younger tumor patients were five times higher than that in healthy senior individuals (Wilcoxon rank-sum test $p = 0.004$, **Supplementary Figure S1C**). The MMR-related MS in CAN also showed a higher contribution than HI and UC (HI vs. CAN, Wilcoxon rank-sum test $p = 0.046$; UC vs. CAN, Wilcoxon rank-sum test $p = 0.007$, **Figure 1C**). These results suggested that the underlying specific mutagenic processes drove the mutations in CAN, which differed from HI and UC. To further validate this observation, we identified MS from 35 normal tissues of the brain ($n = 9$), colon ($n = 13$), kidney ($n = 13$) (Hoang et al., 2016). Results showed that the identified MS had low similarity with any known MS in the Catalogue of somatic mutations in cancer (COSMIC) database (cosine similarity < 0.75). Although somatic mutations were detected in nearly all normal samples, even with some mutations located on cancer driver genes, we did not find any known MS associated with tumor initiation in the whole-exome data of normal tissue (**Supplementary Figures S1E,F**). These results suggested that most mutations in normal tissues accumulated passively and randomly, without clear evidence of external pathogenic mutagenic processes. Therefore, our results indicated that MS possessed the potential to distinguish cancer patients from inflammation patients and healthy individuals.

Identification of the Cancer-Type-Specific Mutational Signatures Patterns

We next attempted to evaluate the cancer-type-specificity of MS patterns. We collected two independent datasets with ten primary cancer types, including non-small cell lung cancer (NSCLC), ovarian serous cystadenocarcinoma (OV), bladder cancer (BLCA), breast cancer (BRCA), liver cancers (LIHC), stomach adenocarcinoma (STAD), esophageal carcinoma (ESCA), colon adenocarcinoma (COAD), pancreatic cancer (PAAD), and prostate cancer (PRAD) (**Supplementary Table S1**). Results showed that tumor samples from the same tissue origins had a high degree of homogeneity in MS between two independent datasets (**Figure 1D**). In addition to PAAD and PRAD, the MS of other cancer types had been maintained in a stable state (similarity > 0.95). Although the MS of PAAD and PRAD had a slight inconsistency in the two datasets, the similarity of tumors from the same tissue origin was still greater than 0.9. These results suggested that although driver mutations among different individuals were highly diverse, the mutagenic processes in specific cancer types were consistent. Therefore, it was reasonably speculated that MS was a stable and informative tissue-specific molecular biomarker to distinguish cancer types.

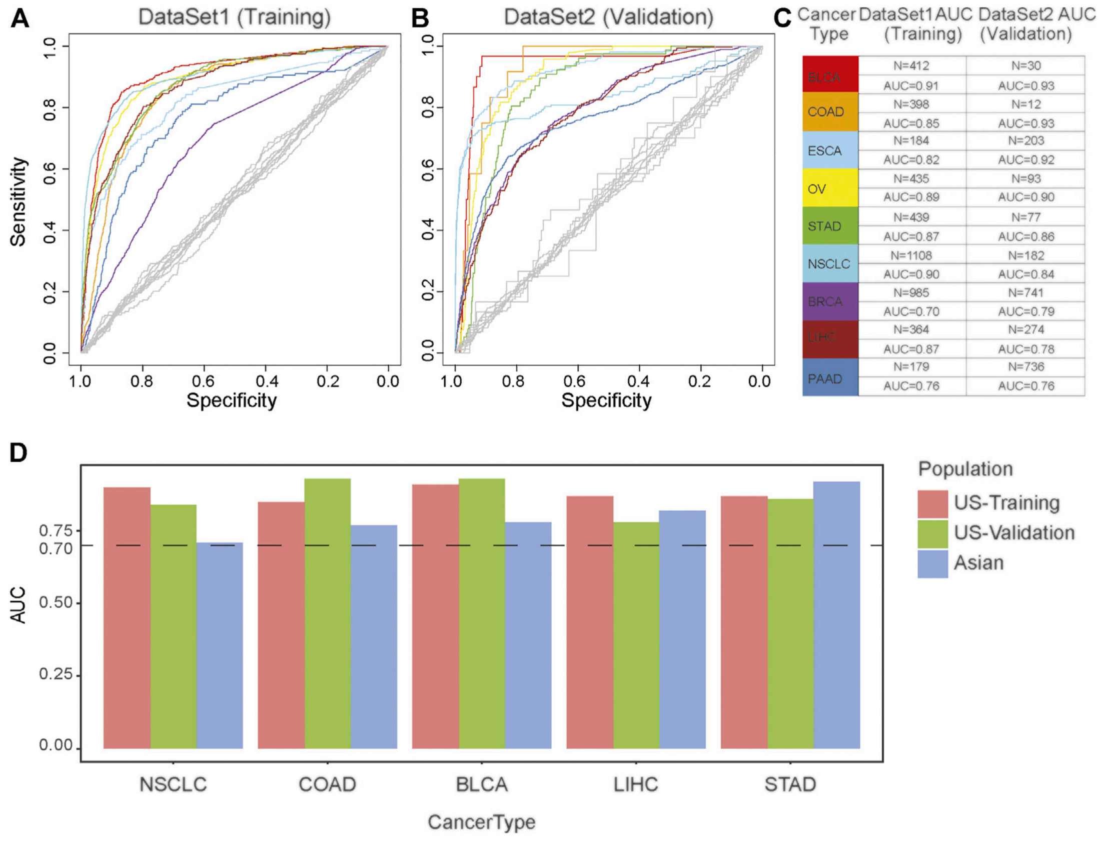
To characterize the landscape of MS in cancers, we identified cancer-type-specific MS (CTS-MS) patterns from The Cancer Genome Atlas (TCGA) dataset (DataSet1). The result showed that the contribution of signatures across different cancer types was distinct (**Figure 1E**; **Supplementary Figure S2**). Specifically, NSCLC highlighted smoking signature, which was previously found in multiple types of lung cancers with probable etiology of tobacco carcinogens (Pfeifer 2010). OV harbored signature associated with the BRCA1 and BRCA2 mutation (Yang et al., 2018). The most common MS in BLCA was related to the misdirected activity of APOBEC3 cytidine deaminases, especially APOBEC3A or APOBEC3B (Robertson et al., 2018). APOBEC related signature and BRCA-mutation-related signature were the main mechanisms of mutations in BRCA. The risk of skin cancer was associated with UV light exposure (Pham et al., 2020). Signatures related to aflatoxin and aristolochic acid were observed in LIHC (Li et al., 2020; Lu et al., 2020; Zhang et al., 2020). STAD and ESCA were enriched in MMR (Meier et al., 2019; Li et al., 2020). The difference in genomic fingerprints between STAD and ESCA was Signature.16, which currently had no clear exposure factor (Wei et al., 2021). The mutations in COAD resulted from Signature.1, which was associated with an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine (Pandey et al., 2019). In summary, our results indicated that CTS-MS implied the origin of the tumors and could be possibly used to detect and localize the cancers.

Mutational Signatures-Based Machine Learning Model for Sensitive Primary Tumor Detection and Classification

To evaluate the performance of MS in cancer diagnosis, we developed a predictive model for each cancer type based on the TCGA databases, including BLCA, COAD, ESCA, OV, STAD, NSCLC, BRCA, LIHC, PAAD, and PRAD. We incorporated the above CTS-MS patterns into a logistic regression algorithm to propose a diagnosis model for each tumor type (**Figure 2A**). We further applied the classifier to predict the tissue of origin in an independent validation dataset with 2,580 additional samples (**Supplementary Tables S1, S2**). The classifier achieved an accurate classification decision, in which the area under the curve (AUC) ranged from 76 to 93% in different cancer types (**Figures 2B,C**). The AUC was relatively higher in cancer types with distinctive MS, such as BLCA (93%), COAD (92.5%), and ESAD (92.5%). However, PRAD was confused with other tumors, possibly due to the lack of specific MS patterns (**Supplementary Figure S3A**). Furthermore, we divided our validation dataset into three groups, including young, middle-aged, and elder samples. Results showed that the performance of our model remained stable across different age groups (**Supplementary Figures S3B–D**). To evaluate the efficacy of MS in inferring primary tumor sites across different populations, we validated our model in an Asian population. We found that the AUC of our model in predicting the tumor origin of the Asian population was higher than 0.7, indicating that our model is stable in different populations (**Figure 2D**). Thus, the above results suggested that CTS-MS were robust candidate biomarkers for the differential diagnosis of various cancer types.

Mutational Signature Patterns of Primary Cancers Maintain in Metastatic Sites

Identification of the primary location of metastatic tumors is essential for precision treatment. To further evaluate the ability of MS to trace tumor location, we performed principal component analysis (PCA) on matched primary and metastatic cancers from 89 lesions (20 patients), including 30 pancreatic cancer and 59 lung cancers (**Supplementary Table S3**). We found that the samples were clustered by tumor origins (**Figure 3A**; **Supplementary Figure S4**). This result was consistent with the study from Connor et al., who found that the MS patterns between primary and metastatic tumors were similar (Connor et al., 2017). Furthermore, different tumor sites from the same individual also showed the same MS pattern (**Figure 3B**; **Supplementary Figure S5**). We compared the MS patterns in matched primary and metastatic cancers and observed high MS consistency between primary cancers and paired metastatic lesions (normalization score > 0.95, **Figures 3A,B**). However, the discrimination efficiency based on the original mutation



Cancer Type	DataSet1 AUC (Training)	DataSet2 AUC (Validation)

BLCA	N=412
	AUC=0.91
		N=30
		AUC=0.93

COAD	N=398
	AUC=0.85
		N=12
		AUC=0.93

ESCA	N=184
	AUC=0.82
		N=203
		AUC=0.92

OV	N=435
	AUC=0.89
		N=93
		AUC=0.90

STAD	N=439
	AUC=0.87
		N=77
		AUC=0.86

NSCLC	N=1108
	AUC=0.90
		N=182
		AUC=0.84

BRCA	N=985
	AUC=0.70
		N=741
		AUC=0.79

LIHC	N=364
	AUC=0.87
		N=274
		AUC=0.78

PAAD	N=179
	AUC=0.76
		N=736
		AUC=0.76

FIGURE 2 | The effectiveness of the cancer diagnosis model based on the MS of the primary tumor. **(A,B)** AUC-curve of cancer diagnosis models in both training **(A)** and validation **(B)** cohort. Random classifiers, indicating the classification accuracies obtained by chance, are shown in gray. **(C)** The value of AUC and the number of patients in both training (left) and validation (right) cohort. **(D)** The model performance across different populations. The vertical axis is the AUC of model. The horizontal axis represents tumor type. Red represents European and American population in the training dataset; Green represents European and American population in the validation dataset; Blue represents Asian populations.

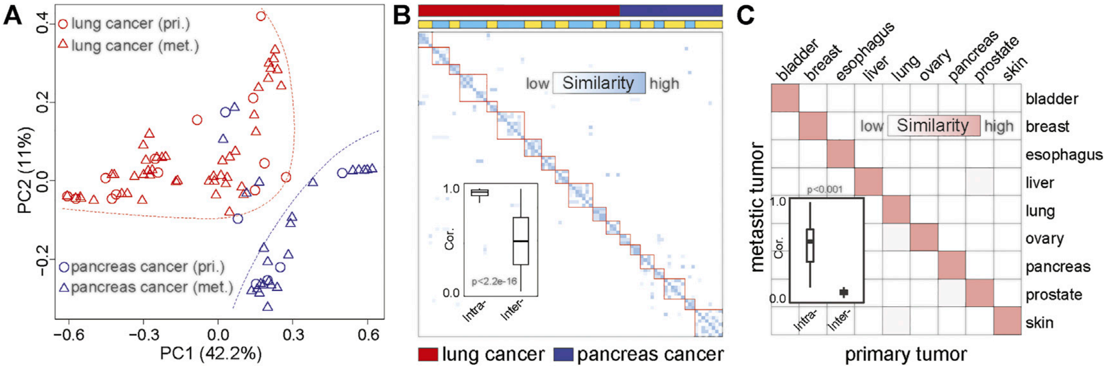
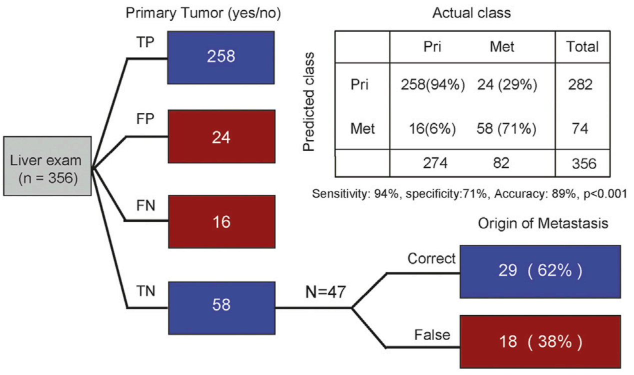


FIGURE 3 | The similarity between metastatic and primary tumors based on MS. **(A)** PCA based on the MS of matched primary and metastatic cancers. The red dots represent the primary lung cancer, the red triangles represent metastatic lung cancer, the blue dots represent the primary pancreatic cancer, and the blue triangles represent metastatic pancreatic cancer. The red dotted line indicates the distribution area of lung cancer, and the blue dotted line indicates the distribution area of pancreatic cancer. pri., primary cancer; met., metastatic cancer. **(B)** The similarity between primary cancer and metastatic cancer based on MS. The darker the color, the higher the similarity. The first line indicates the tumor type. Red represents lung cancer, and blue represents pancreatic cancer. The second line shows the origin of the sample. The same color indicates that the sample is from the same patient. **(C)** The correlation between primary and metastatic cancer based on MS in common cancer. The darker the color, the higher the similarity. The boxplot shows the similarity between the primary and metastatic tumors of the same tumor type and the similarity between the primary and metastatic tumors among different tumor types.



Predicted class	Actual class: Pri	Met	Total
Pri	258(94%)	24 (29%)	282
Met	16(6%)	58 (71%)	74
	274	82	356

FIGURE 4 | Tracing the origin of metastatic tumor based on MS. The first column distinguishes whether it is primary liver cancer. The second column traces the origin of metastatic cancer. TP, true positive; FP, false positive; FN, false negative; TN, true negative. Indicated are sample numbers and detection rates in percentages.

spectrum was lower than that of MS, suggesting that MS can reveal the tissue origin of tumors more effectively than somatic mutations (**Supplementary Figure S6**).

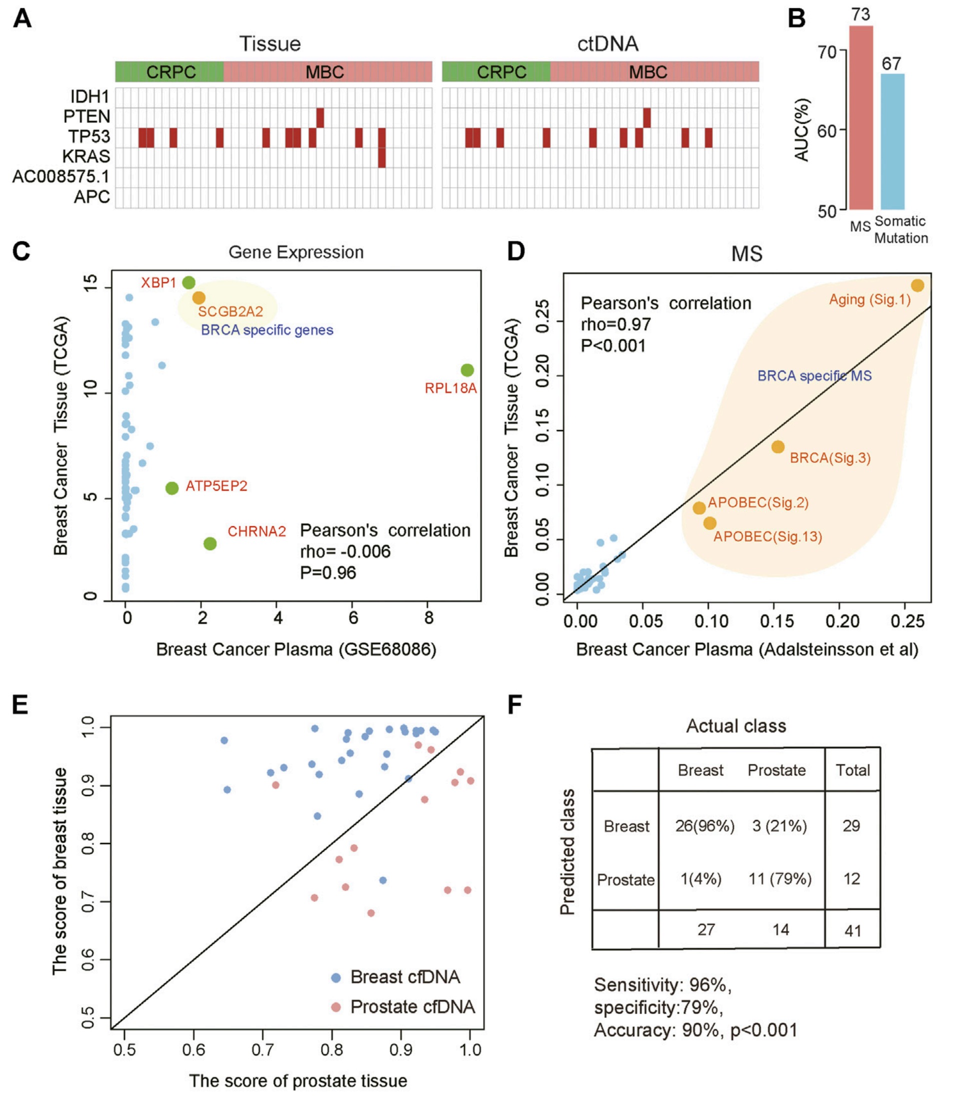
To further validate the similarities between MS across the primary and metastatic tumors, we collected whole-exome data of primary and metastatic tumors of nine cancer types from the previous study (Zhao et al., 2016). We systematically analyzed the homogeneity between metastatic and primary cancer among nine cancer types. As shown in **Figure 3C**, high MS similarities were observed in the primary and metastatic tumor from the same tissue-of-origin (similarity > 0.9), which was significantly higher than the similarity among different cancer types (Wilcoxon rank-sum test p < 0.01). Therefore, our result revealed the high homogeneity of MS among the metastases and primary cancers from the same tissue, indicating that MS was a potential molecular marker for tracing the tissue of origin for metastatic cancers.

Cancer-Type-Specific-Mutational Signatures can Help Identify the Tissue of Origin for Metastatic Cancers

According to the above results, we next sought to evaluate whether CTS-MS was a stable and effective molecular marker for predicting the tissue origin of metastatic cancers. Liver is the most common site of distant metastasis in solid tumors (Riihimaki et al., 2016; Dasari et al., 2017). There is a pressing need for accurate tracing of original tissues (Varghese et al., 2017). We validated the ability of the CTS-MS to identify the tissue origin for metastatic tumor samples in an independent validation dataset that combined a series of 282 primary liver cancer with 74 liver metastatic tumors originating from other organs, including breast, prostate, and esophagus. Firstly, our model accurately distinguished the primary liver cancer and liver metastasis cancer originating from other organs (accuracy: 89%, sensitivity is 94%, specificity is 71%, **Figure 4**). Then, we determined the origin of cancer metastasized to the liver. We identified the origins of metastases with 62% accuracy, in which 75% of breast cancers were correctly classified. And we predicted esophageal cancer with 67% accuracy. However, we only predicted the origin of prostate cancer with 20% accuracy, probably due to the absence of PRAD-specific CTS-MS (**Figure 1E**; **Supplementary Figure S3**). Therefore, these results demonstrated that CTS-MS could help identify the tissue of origin for metastatic cancers.

Cancer-Type-Specific-Mutational Signatures Analysis of Plasma cfDNA Enables Cancer Classification

The advent of non-invasive molecular profiling of plasma cell-free DNA (cfDNA) raises the possibility of inferring a suggested diagnosis in cancer screening. To assess the potential of MS for tracing the tumor origin based on plasma samples, we compared the MS patterns between cfDNA and matched breast and prostate tumor biopsies (Adalsteinsson et al., 2017). We found a high concordance of MS patterns between cfDNA and tissue (Spearman correlation, rho = 0.82, p < 0.001). Somatic mutation and gene expression have been used to predict cancer origins (He et al., 2020a; He et al., 2020b). To explore the efficiency of somatic mutation and gene expression in predicting the tumor origin from blood, we also compared the somatic mutation patterns and gene expression patterns between cfDNA and tumor tissue. We used the cancer-type specific genes (IDH1, PTEN, TP53, KRAS, AC008575.1, APC) in TOOme (He et al., 2020b) to evaluate the performance of somatic mutations detected in tissue or ctNDA for identifying the tumor tissue origin. We found that somatic mutations were detected in 26.8% (11/41) of tissue samples using these genes. The performance was even lower in paired ctDNA samples, with only 24.4% (10/41) detection rate (**Figure 5A**). Importantly, these gene mutations cannot distinguish breast cancer from prostate cancer based on these gene mutations. Thus, the above observations indicated that the performance of somatic mutations for inferring cancer tissue-of-origin was limited due to the substantial overlap in mutational profiles across different cancer types. Then, we compared the efficiency of MS and somatic mutations to identify the tumor from ctDNA, based on the somatic mutations detected from plasma of 111 lung cancer patients and 78 benign lung nodules patients (Chen et al., 2021). We found that MS was able to distinguish tumor from non-tumor patients better than mutations (AUC:0.73 vs. 0.67, **Figure 5B**). Next, we compared the expression similarity between tissues and plasma from breast cancer patients based on the genes used in TOOme. Our results indicated that the gene expression pattern differed between tissue and plasma of breast cancer. Almost all genes used to infer tumor tissue origin in TOOme were not expressed in plasma (Pearson correlation: −0.006, p = 0.96, **Figure 5C**). However, breast cancer-specific MS could be detected from ctDNA (**Figure 5D**). These analyses showed that MS is a reliable and stable biomarker for predicting the tumor tissue origin from plasma, compared with somatic



	Breast	Prostate	Total
Breast	26(96%)	3 (21%)	29
Prostate	1(4%)	11 (79%)	12
	27	14	41

Sensitivity: 96%,
specificity:79%,
Accuracy: 90%, p<0.001

FIGURE 5 | Distinguishing different cancer types based on the MS patterns and somatic mutations called from plasma ctDNA data. **(A)** The specific mutations in tissue (left) and ctDNA (right). Red indicates that mutation was detected. CRPC: prostate cancer; MBC: metastatic breast cancer. **(B)** The efficiency of MS and somatic mutations to distinguish lung cancer patients from benign lung nodules patients from ctDNA data. **(C,D)** The correlation between plasma and tissue in breast cancer based on gene expression **(C)** and MS **(D)**. Orange indicates breast cancer-specific markers. **(E,F)** Combined MS patterns and somatic mutations called from plasma ctDNA data distinguish breast cancer and prostate cancer. The red points represent prostate cancer and the blue points represent breast cancer. The horizontal axis represents the score of the prostate cancer model and the vertical axis represents the score of the breast cancer model.

mutation and RNA expression. Then, our model was further used to distinguish breast and prostate cancers based on MS patterns of cfDNA and achieved 71% accuracy. However, the model based on the mutation spectrum called cfDNA cannot distinguish these two tumor types (**Supplementary Figure S7**). We integrated the mutation profile of cfDNA and MS to build diagnosis models. The results showed that the performance of these diagnosis models had been significantly optimized. We predicted the tissue origin with 90% accuracy (sensitivity is 96%, specificity is 79%, **Figures 5E,F**). In summary, our analysis proved that the combination of MS and mutational profile was an available method to detect and localize cancers from peripheral blood.

DISCUSSION

Using the whole-exome sequencing data from tumors and cfDNA, we demonstrated that MS pattern was a potential approach for tumor detection and localization with high accuracy and robustness. First, we found that the somatic mutations in healthy individuals and inflammation patients were not associated with any known tumor initiation-related MS in the COSMIC database. This observation indicated that MS might separate healthy/inflammation patients and tumor patients. To further investigate whether MS could distinguish different tumor types, we analyzed the MS landscape of tumors from TCGA. Our results showed that different cancer types had specific MS patterns and validated this result in an independent dataset.

Moreover, using the CTS-MS, we could predict the tumor origin with high accuracy among primary and metastatic cancer. Notably, MS could better distinguish cancers from different tissues than somatic mutations. Finally, integrating the mutation profile and MS identified from cfDNA, we could predict the tissue origin of tumors with high accuracy. Therefore, our study showed that MS was a robust molecular marker for cancer diagnosis.

Lines of evidence indicate that the human body accumulates random mutations with age (Blokzijl et al., 2016; Hoang et al., 2016; Lodato et al., 2018; Zhang et al., 2019). The inflammation states accelerate this accumulation, such as ulcerative colitis, inflammatory bowel, or cirrhosis diseases (Brunner et al., 2019; Moore et al., 2020; Olafsson et al., 2020). The critical question is whether these accumulations of the somatic mutation have a functional impact or increased cancer risk. Our results indicated that the somatic mutations in healthy individuals had no functional relevance. In contrast, somatic mutations in tumor patients were functionally clustered and were related to specific biological processes, such as DNA damage repair deficiency. Our study showed that MS could distinguish between healthy individuals and tumor patients.

Some previous studies have reported that diverse ethnic populations have different mutational landscapes in the same type of cancer (Yao et al., 2016; Jia et al., 2017). However, the MS-based tumor tracing model in our study showed comparable performance between Asian and European and American populations for most of the tumor types, such as liver cancer, non-small cell lung cancer, and bladder cancer. This observation indicated that MS was a stable marker for predicting the tumor tissue origin in different populations. Consistently, Zhang et al. reported that MS patterns were shared in different populations with liver cancer, including Signature.5, Signature.22, and Signature.24 (Zhang et al., 2017; Zhang and Guan, 2021).

Notably, with one or more confirmed metastatic malignant lesions but the undetectable primary origin, cancers of unknown primary (CUP) make up 3–5% of total cancer diagnoses and have a very poor prognosis with a median survival of 6–16 months (Varadhachary and Raber, 2014; Conway et al., 2019). Refining the diagnostic classification of CUP patients can facilitate the selection of potentially effective therapies (Varghese et al., 2017). We found that the MS of the primary and metastatic cancers from identical tissue were highly consistent in whole-exome sequencing, indicating the tumor traceability of MS for metastatic cancers. We distinguished the malignant liver lesions originating from other tissues and primary liver tumors with high accuracy, indicating that our MS-based model could trace the origin of the metastatic tumor. Besides, MS inferred from cfDNA was highly compatible with tumor biopsies. Since liquid biopsy is increasingly used for cancer screening and diagnosis, our method may help infer the tissue origin by cfDNA detection.

In this study, although we demonstrated the potential diagnostic value of MS in determining the cancer origin by two independent datasets, more samples needed to be included to train more robust and precise models. Besides, only a limited number of MS have been discovered in the human tissue. The etiology and exposure factors of the majority of MS remain unclear currently (Alexandrov et al., 2013). With the development of sequencing technology, more reliable cancer-related MS will be determined, allowing more features could be included in our model to achieve higher accuracy.

In conclusion, we showed that MS was a reliable biomarker for tumor detection and localization. Our study will provide vital information for clinical diagnosis and tracing tumor origin for cancers without known primary sites.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

QW, RG, WWa, and KL had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. ZW, TZ, WWu, and LW contributed equally. Concept and design: QW, RG, WWa, and KL. Data collection: JL, BH, and YuL. Data analysis and interpretation: ZW, TZ, WWu, LW, YaL, and PL.

FUNDING

This work was supported by the grants from the National Natural Science Foundation of China (Grant Nos. 81972358, 91959113, 81972188), Basic Research Program of Jiangsu Province (Grant No. BK20180036), Natural Science Foundation of Jiangsu Province (Grant No. BK20210530), Medical Important Talents of Jiangsu Province (ZDRCA2016024), China Postdoctoral Science Foundation (Grant No. 2021M691645).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.883791/full#supplementary-material

REFERENCES

Adalsteinsson, V. A., Ha, G., Freeman, S. S., Choudhury, A. D., Stover, D. G., Parsons, H. A., et al. (2017). Scalable Whole-Exome Sequencing of Cell-Free DNA Reveals High Concordance with Metastatic Tumors. *Nat. Commun.* 8, 1324. doi:10.1038/s41467-017-00965-y

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., et al. (2016). Mutational Signatures Associated with Tobacco Smoking in Human Cancer. *Science* 354, 618–622. doi:10.1126/science.aag0299

Alexandrov, L. B., Nik-Zainal, S., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., et al. (2013). Signatures of Mutational Processes in Human Cancer. *Nature* 500, 415–421. doi:10.1038/nature12477

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., et al. (2016). Tissue-Specific Mutation Accumulation in Human Adult Stem Cells during Life. *Nature* 538, 260–264. doi:10.1038/nature19768

Brunner, S. F., Roberts, N. D., Wylie, L. A., Moore, L., Aitken, S. J., Davies, S. E., et al. (2019). Somatic Mutations and Clonal Dynamics in Healthy and Cirrhotic Human Liver. *Nature* 574, 538–542. doi:10.1038/s41586-019-1670-9

Chen, K. Z., Sun, J. L., Zhao, H., Jiang, R. J. F., Zheng, J. C., Li, Z. L., et al. (2021). Non-Invasive Lung Cancer Diagnosis and Prognosis Based on Multi-Analyte Liquid Biopsy. *Mol. Cancer* 20 (1), 23. doi:10.1186/s12943-021-01323-9

Connor, A. A., Denroche, R. E., Jang, G. H., Timms, L., Kalimuthu, S. N., Selander, I., et al. (2017). Association of Distinct Mutational Signatures with Correlates of Increased Immune Activity in Pancreatic Ductal Adenocarcinoma. *JAMA Oncol.* 3, 774–783. doi:10.1001/jamaoncol.2016.3916

Conway, A.-M., Mitchell, C., Kilgour, E., Brady, G., Dive, C., and Cook, N. (2019). Molecular Characterisation and Liquid Biomarkers in Carcinoma of Unknown Primary (CUP): Taking the 'U' Out of 'CUP'. *Br. J. Cancer* 120, 141–153. doi:10.1038/s41416-018-0332-2

Dasari, A., Shen, C., Halperin, D., Zhao, B., Zhou, S., Xu, Y., et al. (2017). Trends in the Incidence, Prevalence, and Survival Outcomes in Patients with Neuroendocrine Tumors in the United States. *JAMA Oncol.* 3, 1335–1342. doi:10.1001/jamaoncol.2017.0589

Degasperi, A., Amarante, T. D., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D., et al. (2020). A Practical Framework and Online Tool for Mutational Signature Analyses Show Intertissue Variation and Driver Dependencies. *Nat. Cancer* 1, 249–263. doi:10.1038/s43018-020-0027-5

Dietlein, F., and Eschner, W. (2014). Inferring Primary Tumor Sites from Mutation Spectra: A Meta-Analysis of Histology-Specific Aberrations in Cancer-Derived Cell Lines. *Hum. Mol. Genet.* 23, 1527–1537. doi:10.1093/hmg/ddt539

Fu, Z., Chen, X., Yang, X., and Li, Q. (2019). Diagnosis of Primary Clear Cell Carcinoma of the Vagina by 18F-FDG PET/CT. *Clin. Nucl. Med.* 44, 332–333. doi:10.1097/rlu.0000000000002463

Gehring, J. S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: Inferring Mutational Signatures from Single-Nucleotide Variants. *Bioinformatics* 31, 3673–3675. doi:10.1093/bioinformatics/btv408

He, B. S., Dai, C., Lang, J. D., Bing, P. P., Tian, G., Wang, B., et al. (2020a). A Machine Learning Framework to Trace Tumor Tissue-Of-Origin of 13 Types of Cancer Based on DNA Somatic Mutation. *Bba-Mol Basis Dis.* 1866 (11), 165916. doi:10.1016/j.bbadis.2020.165916

He, B. S., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020b). TOOme: A Novel Computational Framework to Infer Cancer Tissue-Of-Origin by Integrating Both Gene Mutation and Expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms Underlying Mutational Signatures in Human Cancers. *Nat. Rev. Genet.* 15, 585–598. doi:10.1038/nrg3729

Hoang, M. L., Kinde, I., Tomasetti, C., McMahon, K. W., Rosenquist, T. A., Grollman, A. P., et al. (2016). Genome-Wide Quantification of Rare Somatic Mutations in normal Human Tissues Using Massively Parallel Sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 113, 9846–9851. doi:10.1073/pnas.1607794113

Jerjes, W., Upile, T., Petrie, A., Riskalla, A., Hamdoon, Z., Vourvachis, M., et al. (2010). Clinicopathological Parameters, Recurrence, Locoregional and Distant Metastasis in 115 T1-T2 Oral Squamous Cell Carcinoma Patients. *Head Neck Oncol.* 2, 9. doi:10.1186/1758-3284-2-9

Jia, F., Teer, J. K., Knepper, T. C., Lee, J. K., Zhou, H.-H., He, Y.-J., et al. (2017). Discordance of Somatic Mutations between Asian and Caucasian Patient Populations with Gastric Cancer. *Mol. Diagn. Ther.* 21, 179–185. doi:10.1007/s40291-016-0250-z

Jiao, W., Polak, P., Atwal, G., Polak, P., Karlic, R., Cuppen, E., et al. (2020). A Deep Learning System Accurately Classifies Primary and Metastatic Cancers Using Passenger Mutation Patterns. *Nat. Commun.* 11, 728. doi:10.1038/s41467-019-13825-8

Jurmeister, P., Bockmayr, M., Seegerer, P., Bockmayr, T., Treue, D., Montavon, G., et al. (2019). Machine Learning Analysis of DNA Methylation Profiles Distinguishes Primary Lung Squamous Cell Carcinomas from Head and Neck Metastases. *Sci. Transl Med.* 11 (509), eaaw8513. doi:10.1126/scitranslmed.aaw8513

Li, L., Feng, Q., and Wang, X. (2020). PreMSIm: An R Package for Predicting Microsatellite Instability from the Expression Profiling of a Gene Panel in Cancer. *Comput. Struct. Biotechnol. J.* 18, 668–675. doi:10.1016/j.csbj.2020.03.007

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., et al. (2018). Aging and Neurodegeneration are Associated with Increased Mutations in Single Human Neurons. *Science* 359, 555–559. doi:10.1126/science.aao4426

Lu, Z. N., Luo, Q., Zhao, L. N., Shi, Y., Wang, N., Wang, L., et al. (2020). The Mutational Features of Aristolochic Acid-Induced Mouse and Human Liver Cancers. *Hepatology* 71, 929–942. doi:10.1002/hep.30863

Manca, A., Paliogiannis, P., Colombino, M., Casula, M., Lissia, A., Botti, G., et al. (2019). Mutational Concordance between Primary and Metastatic Melanoma: A Next-Generation Sequencing Approach. *J. Transl Med.* 17, 289. doi:10.1186/s12967-019-2039-4

Marquard, A. M., Birkbak, N. J., Thomas, C. E., Favero, F., Krzystanek, M., Lefebvre, C., et al. (2015). TumorTracer: A Method to Identify the Tissue of Origin from the Somatic Mutations of a Tumor Specimen. *BMC Med. Genomics* 8, 58. doi:10.1186/s12920-015-0130-0

Martincorena, I., and Campbell, P. J. (2016). Somatic Mutation in Cancer and normal Cells. *Science* 353, 132. doi:10.1126/science.aab4082

Meier, B., Volkova, N. V., Hong, Y., Schofield, P., Campbell, P. J., Gerstung, M., et al. (2019). Mutational Signatures of DNA Mismatch Repair Deficiency in *C. E* and Human Cancers. *Genome Res.* 29, 1566. doi:10.1101/gr.255596.119

Moore, L., Leongamornlert, D., Coorens, T. H. H., Sanders, M. A., Ellis, P., Dentro, S. C., et al. (2020). The Mutational Landscape of normal Human Endometrial Epithelium. *Nature* 580, 640–646. doi:10.1038/s41586-020-2214-z

Nanki, K., Fujii, M., Shimokawa, M., Matano, M., Nishikori, S., Date, S., et al. (2020). Somatic Inflammatory Gene Mutations in Human Ulcerative Colitis Epithelium. *Nature* 577, 254–259. doi:10.1038/s41586-019-1844-5

Olafsson, S., McIntyre, R. E., Coorens, T., Butler, T., Jung, H., Robinson, P. S., et al. (2020). Somatic Evolution in Non-Neoplastic IBD-Affected Colon. *Cell* 182, 672–684. doi:10.1016/j.cell.2020.06.036

Pandey, P., Yang, Z., Shibata, D., Marjoram, P., and Siegmund, K. D. (2019). Mutational Signatures in Colon Cancer. *BMC Res. Notes* 12, 788. doi:10.1186/s13104-019-4820-0

Pfeifer, G. P. (2010). Environmental Exposures and Mutational Patterns of Cancer Genomes. *Genome Med.* 2, 54. doi:10.1186/gm175

Pham, T. V., Boichard, A., Goodman, A., Riviere, P., Yeerna, H., Tamayo, P., et al. (2020). Role of Ultraviolet Mutational Signature versus Tumor Mutation burden in Predicting Response to Immunotherapy. *Mol. Oncol.* 14, 1680–1694. doi:10.1002/1878-0261.12748

Pilati, C., Shinde, J., Alexandrov, L. B., Assié, G., André, T., Hélias-Rodzewicz, Z., et al. (2017). Mutational Signature Analysis Identifies MUTYH Deficiency in Colorectal Cancers and Adrenocortical Carcinomas. *J. Pathol.* 242, 10–15. doi:10.1002/path.4880

Riihimäki, M., Hemminki, A., Sundquist, K., Sundquist, J., and Hemminki, K. (2016). The Epidemiology of Metastases in Neuroendocrine Tumors. *Int. J. Cancer* 139, 2679–2686. doi:10.1002/ijc.30400

Roberts, S. A., and Gordenin, D. A. (2014). Hypermutation in Human Cancer Genomes: Footprints and Mechanisms. *Nat. Rev. Cancer* 14, 786–800. doi:10.1038/nrc3816

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., et al. (2018). Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* 174, 1033. doi:10.1016/j.cell.2018.07.036

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., and Swanton, C. (2016). DeconstructSigs: Delineating Mutational Processes in Single Tumors

Distinguishes DNA Repair Deficiencies and Patterns of Carcinoma Evolution. *Genome Biol.* 17, 31. doi:10.1186/s13059-016-0893-4

Saudemont, P., Quanico, J., Robin, Y.-M., Baud, A., Balog, J., Fatou, B., et al. (2018). Real-Time Molecular Diagnosis of Tumors Using Water-Assisted Laser Desorption/Ionization Mass Spectrometry Technology. *Cancer Cell* 34, 840–851. doi:10.1016/j.ccell.2018.09.009

Thomson, P. J. (2018). Perspectives on Oral Squamous Cell Carcinoma Prevention-Proliferation, Position, Progression and Prediction. *J. Oral Pathol. Med.* 47, 803–807. doi:10.1111/jop.12733

Varadhachary, G. R., and Raber, M. N. (2014). Cancer of Unknown Primary Site. *N. Engl. J. Med.* 371, 757–765. doi:10.1056/nejmra1303917

Varghese, A. M., Arora, A., Capanu, M., Camacho, N., Won, H. H., Zehir, A., et al. (2017). Clinical and Molecular Characterization of Patients with Cancer of Unknown Primary in the Modern Era. *Ann. Oncol.* 28, 3015–3021. doi:10.1093/annonc/mdx545

Wei, R., Li, P., He, F., Wei, G., Zhou, Z., Su, Z., et al. (2021). Comprehensive Analysis Reveals Distinct Mutational Signature and its Mechanistic Insights of Alcohol Consumption in Human Cancers. *Brief Bioinform* 22 (3), bbaa066. doi:10.1093/bib/bbaa066

Welch, J. S., Ley, T. J., Link, D. C., Miller, C. A., Larson, D. E., Koboldt, D. C., et al. (2012). The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* 150, 264–278. doi:10.1016/j.cell.2012.06.023

Yang, S. Y. C., Lheureux, S., Karakasis, K., Burnier, J. V., Bruce, J. P., Clouthier, D. L., et al. (2018). Landscape of Genomic Alterations in High-Grade Serous Ovarian Cancer from Exceptional Long- and Short-Term Survivors. *Genome Med.* 10, 81. doi:10.1186/s13073-018-0590-x

Yao, S., Johnson, C., Hu, Q., Yan, L., Liu, B., Ambrosone, C. B., et al. (2016). Differences in Somatic Mutation Landscape of Hepatocellular Carcinoma in Asian American and European American Populations. *Oncotarget* 7, 40491–40499. doi:10.18632/oncotarget.9636

Zhang, B-F., and Guan, X-Y. (2021). Racial Difference of Mutational Signature in Hepatocellular Carcinoma. *Hepatoma Res.* 7, 62. doi:10.20517/2394-5079.2021.81

Zhang, L., Dong, X., Lee, M., Maslov, A. Y., Wang, T., and Vijg, J. (2019). Single-Cell Whole-Genome Sequencing Reveals the Functional Landscape of Somatic Mutations in B Lymphocytes across the Human Lifespan. *Proc. Natl. Acad. Sci. U.S.A.* 116, 9014–9019. doi:10.1073/pnas.1902510116

Zhang, W., Liu, Y., Liang, B., Zhang, Y., Zhong, X., Luo, X., et al. (2020). Probabilistic Risk Assessment of Dietary Exposure to Aflatoxin B1 in Guangzhou, China. *Sci. Rep.* 10, 7973. doi:10.1038/s41598-020-64295-8

Zhang, W., He, H., Zang, M., Wu, Q., Zhao, H., Lu, L.-l., et al. (2017). Genetic Features of Aflatoxin-Associated Hepatocellular Carcinoma. *Gastroenterology* 153, 249–262. doi:10.1053/j.gastro.2017.03.024

Zhao, Z.-M., Zhao, B., Bai, Y., Iamarino, A., Gaffney, S. G., Schlessinger, J., et al. (2016). Early and Multiple Origins of Metastatic Lineages within Primary Tumors. *Proc. Natl. Acad. Sci. U.S.A.* 113, 2140–2145. doi:10.1073/pnas.1525677113

Zou, X., Owusu, M., Harris, R., Jackson, S. P., Loizou, J. I., and Nik-Zainal, S. (2018). Validating the Concept of Mutational Signatures with Isogenic Cell Models. *Nat. Commun.* 9, 1744. doi:10.1038/s41467-018-04052-8

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.

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.

ORIGINAL RESEARCH
published: 13 May 2022
doi: 10.3389/fbioe.2022.849798

Pan-Cancer Single-Cell Analysis Reveals the Core Factors and Pathway in Specific Cancer Stem Cells of Upper Gastrointestinal Cancer

*Leijie Li[1], Yujia Zhang[1], Yongyong Ren[1], Zhiwei Cheng[1], Yuening Zhang[1], Xinbo Wang[1], Hongyu Zhao[2] and Hui Lu[1]**

[1]SJTU-Yale Joint Center for Biostatistics and Data Science, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [2]Department of Biostatistics, Yale University, New Haven, CT, United States

Upper gastrointestinal cancer (UGIC) is an aggressive carcinoma with increasing incidence and poor outcomes worldwide. Here, we collected 39,057 cells, and they were annotated into nine cell types. By clustering cancer stem cells (CSCs), we discovered the ubiquitous existence of sub-cluster CSCs in all UGICs, which is named upper gastrointestinal cancer stem cells (UGCSCs). The identification of UGCSC function is coincident with the carcinogen of UGICs. We compared the UGCSC expression profile with 215,291 single cells from six other cancers and discovered that UGCSCs are specific tumor stem cells in UGIC. Exploration of the expression network indicated that inflammatory genes (*CXCL8*, *CXCL3*, *PIGR*, and *RNASE1*) and Wnt pathway genes (*GAST*, *REG1A*, *TFF3*, and *ZG16B*) are upregulated in tumor stem cells of UGICs. These results suggest a new mechanism for carcinogenesis in UGIC: mucosa damage and repair caused by poor eating habits lead to chronic inflammation, and the persistent chronic inflammation triggers the Wnt pathway; ultimately, this process induces UGICs. These findings establish the core signal pathway that connects poor eating habits and UGIC. Our system provides deeper insights into UGIC carcinogens and a platform to promote gastrointestinal cancer diagnosis and therapy.

Keywords: upper gastrointestinal cancers, cancer stem cell, single-cell sequence data, pan-cancer analysis, oncogene

OPEN ACCESS

Edited by:
Zhaoyuan Fang,
Zhejiang University, China

Reviewed by:
Binhua Tang,
Hohai University, China
Subhayan Sur,
Saint Louis University, United States

****Correspondence:***
Hui Lu
huilu@sjtu.edu.cn

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *06 January 2022*
Accepted: *07 April 2022*
Published: *13 May 2022*

Citation:
Li L, Zhang Y, Ren Y, Cheng Z, Zhang Y, Wang X, Zhao H and Lu H (2022) Pan-Cancer Single-Cell Analysis Reveals the Core Factors and Pathway in Specific Cancer Stem Cells of Upper Gastrointestinal Cancer. Front. Bioeng. Biotechnol. 10:849798. doi: 10.3389/fbioe.2022.849798

INTRODUCTION

Upper gastrointestinal cancer (UGIC), including head and neck squamous cell carcinoma (HNSCC), esophageal cancer (EC), and gastric cancer (GC), is one of the malignant tumors that seriously threaten the human health (Yamada et al., 2011). Its occurrence is mainly associated with unhealthy eating habits and lifestyle and their consequences, including low intake of fruits and vegetables (Akhtar, 2013), smoking (Gandini et al., 2008), drinking (Goldstein et al., 2010; Zhang et al., 2012; González et al., 2013), and high body mass index (BMI). The global incidence of UGIC has significantly increased in recent years (Bray et al., 2018). Patients with UGIC account for a large proportion of all patients with malignant tumors (Sung et al., 2021). UGIC has a poorer prognosis and lower overall survival rate than other cancers (Sung et al., 2021). GC is the fifth most prevalent cancer and the third leading death cause of patients with cancers on a global scale (Yin et al., 2020). The 5-year survival rate of patients with EC is not more than 20% worldwide (Zhang, 2013). Because of the increasing incidence, the high relapse and metastasis rate, and the low overall survival rate, studies on the molecular mechanism of UGICs or gastrointestinal pan-cancer are imperative.

In recent years, the growing number of patients has prompted many studies on gastrointestinal tumors (Chakravarthy et al., 2018; Yang et al., 2020; Cui et al., 2021). At present, researchers have discovered many biomarkers for the diagnosis and treatment of gastrointestinal cancer, including human epidermal growth factor receptor2 (HER2) (Li Z. et al., 2020), mismatch repair deficiency/microsatellite instability (dMMR/ MSI-H) (Dhakras et al., 2020), and programmed death-ligand 1 (PD- L1) (Dai et al., 2021). In addition, there are many new biomarkers under investigation, including neurotrophic-tropomyosin receptor kinase (NTRK) (Westphalen et al., 2019), claudin-18 (CLDN18) (Zhang et al., 2020), Rho GTPase-activating protein 26 (ARHGAP26) (Dhakras et al., 2020), fibroblast growth factor receptor (FGFR) (Babina and Turner, 2017), lymphocyte-activation gene 3 (LAG3) (Saleh et al., 2019), and T-cell immunoglobulin and mucin-domain containing-3 (TIM3) (Wang et al., 2017). However, only few clinical trials on UGIC patients have shown positive curative effects; the underlying mechanisms remain elusive so far. Nearly 50% of patients in good conditions will still suffer from local recurrence or systematic metastasis after aggressive treatment (Dhakras et al., 2020; Sung et al., 2021). It seems that most of the works aim at general tumor cells rather than cancer stem cells (CSCs) in UGIC. It is because CSCs are difficult to isolate due to the limitation of early experimental conditions and heterogeneity of CSCs (Clarke et al., 2006; Sreepadmanabh and Toley, 2018). Considering that the digestive tract organs share a common external environment and perform similar functions in a system, diet-induced mucosal lesions may have similar effects on cancer of the mouth, esophagus, and stomach (Haas et al., 2012). Therefore, it is necessary to take oral cancer, esophageal cancer, and gastric cancer as a whole, that is, UGIC, for integration research.

Some laboratories have conducted pan-cancer research on UGICs. Tran et al.'s pan-cancer study on somatic mutations found that leukocyte antigen-restricted T-cell receptors targeted the KRAS (G12D) hotspot driver mutation found in many human gastrointestinal cancers (Tran et al., 2015). Another study observed that IL-6 is the main communication medium for tumor cells and cancer-related fibroblasts in a murine model (Johnson et al., 2018). IL-6 deletion inhibits the occurrence of gastrointestinal tumors through STAT3 and MEK1/2 signals (Karakasheva et al., 2018). Dana–Farber Cancer Institute discovered the new immune checkpoint biomarker TET1 and PD-1 ligands (CD274 and PDCD1LG2) (Thienpont et al., 2016; Bu et al., 2021; Rahman et al., 2021). But fewer studies focus on CSCs. The problems of poor prognosis and a high recurrence rate still require more intensive studies in UGIC.

In this work, to verify the pathogenesis and therapeutic targets of UGIC, we performed the pan-cancer analysis on UGIC. Our results identified the unique CSCs in UGIG, which are named upper gastrointestinal common cancer stem cells (UGCSCs). The core regulation network of UGCSCs suggested that inflammation-related genes, namely, *CXCL8*, *PIGR*, and *CXCL3*, and Wnt pathway-related genes, namely, *GAST*, *REG1A*, *TFF3*, and *ZG16B*, are activated. Further analysis indicated that mucosal damage and inflammation caused by poor dietary habits trigger the Wnt pathway and eventually induce UGIC. In addition, GAST and TFF3 activate phosphatidylinositol 3-kinase (PI3k)/Ras to enhance the metastasis and invasion of UGIC. Taken together, these results pave the way for the better diagnosis and treatment of UGIC.

METHODS

Data Collection and Processing

The data were collected from the published literature (**Table 1**). For different sequencing methods of single-cell data, specific analysis procedures were applied. For Drop-seq single-cell data, Cell Ranger software (Freytag et al., 2018) (3.0.1) was adopted to calculate the cell expression counts. For Smart-seq2 single-cell sequencing data, we operated cell expression matrixes provided in the original article. The expression matrix file was then imported into R 3.6.2 for subsequent analysis.

Data Normalization and Batch Effect Correction

First, we used Seurat (Stuart et al., 2019) (3.1.4) to filter the quality of cells and delete all cells with more than 6000 expressed genes or less than 201 genes. A total of 39,057 UGIC cells and 215,291 other cancer cells were obtained. Next, standardized integration processing was performed on the cell level, sample level, and study level.

Cell Level Standardization

The logarithmic percentage of gene expression in cells was adopted as the standardized integration of data between different cells in the sample (Butler et al., 2018). The value of the expression of gene x in a cell was divided by the value of the expression of all genes in this cell and multiplied by the scale amplification factor, which is set to 10000 in this experiment. Then, the logarithm of this value is the normalized value of the expression of gene x in the cell. This process can reduce the deviation of gene expression values caused by different sequencing depths and sequencing methods. The **formula (1)** is described as follows:

$$x_i^{'} = \log_{10}\left(\frac{x_i}{\sum_{i\in U} x_i} * 10000\right), \quad (1)$$

where x_i represents the expression value of gene i. $x_i^{'}$ represents the expression value of gene i after normalization. U represents the gene set in a certain cell.

Sample Level Standardization

We diminished gene features to avert the dimension disaster problem in the single-cell expression matrix. First, the logarithm of gene expression means and variances was calculated. Next, we

TABLE 1 | Single-cell RNA-seq data of UGIC.

Species	Tumor type	Tissue	Sequence type	Cell number	Sample	PubMed ID
Human	EC	Esophagus	Smart-seq2	366	5	30223068 Wu et al. (2018)
Human	HNSCC	Oral cavity	Smart-seq2	4762	15	29198524 Puram et al. (2017)
Human	Early GC	Stomach	10x Genomics	4110	1	32209487 Zhang et al. (2019a)
Human	GC	Stomach	10x Genomics	29817	9	32532891 Zhang et al. (2021a)

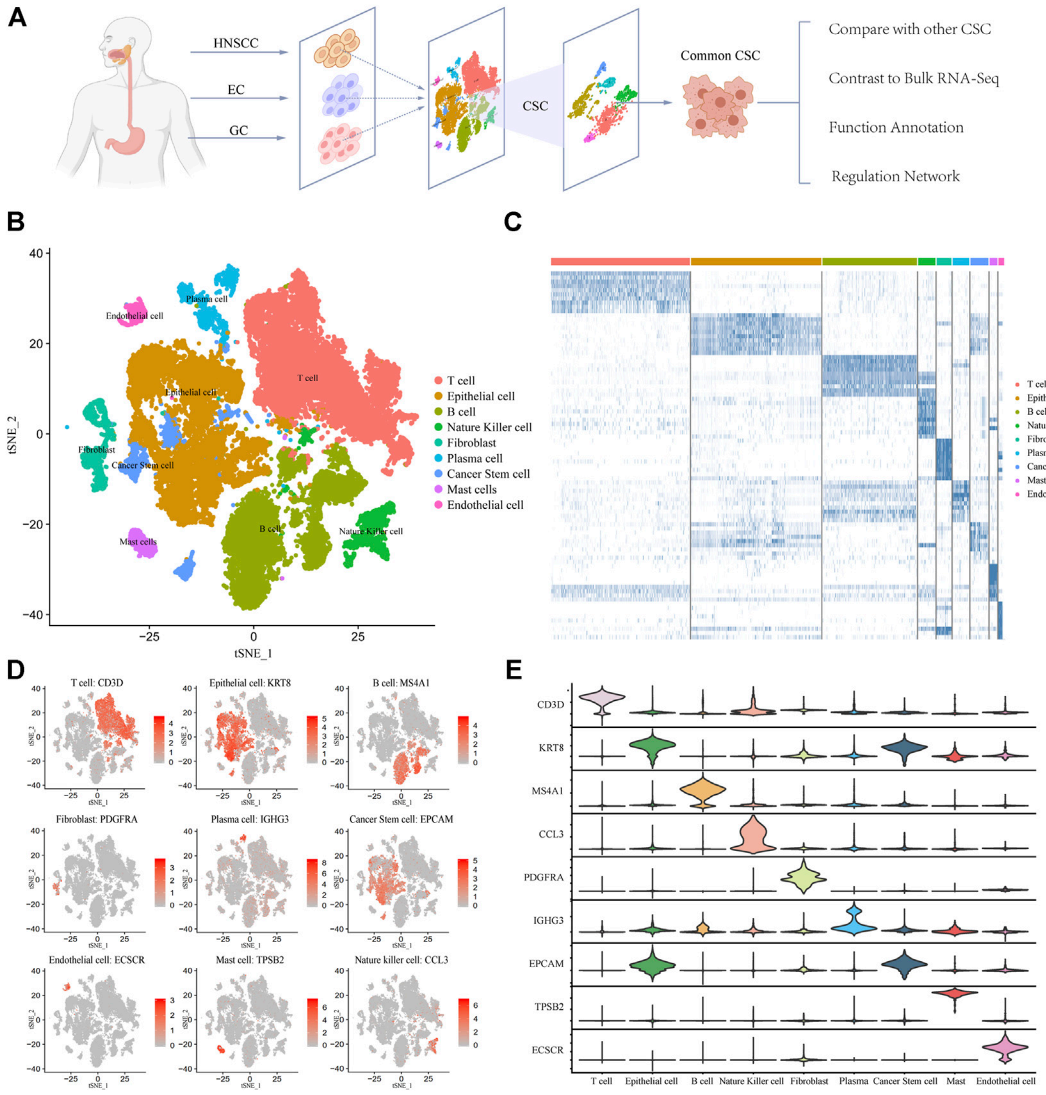


FIGURE 1 | Expression profiling of 39,057 single cells in UGIC. **(A)** Workflow of sample processing, cell type annotation, and functional analysis for 30 samples in UGIC. **(B)** t-SNE of 39,057 cells profiled here, with each cell color-coded for the associated cell type. **(C)** Heatmap of the expression pattern in each cell type. **(D)** Expression of marker genes for the cell types defined above each panel. **(E)** Expression trend of marker genes for each cell type in the violin chart.

fitted a line regression model to represent the relationship between the two values using the local polynomial regression. Next, we normalized the gene expression value through the mean value and expected variance of the model. Finally, the top variable 2000 gene features were selected for the subsequent analysis based on the normalized expression value.

Study Level Standardization

We conducted an integrated analysis of multiple samples, by looking for similar sites between cells. First, the dimensionality was reduced by using canonical correlation analysis (CCA) (Andrew et al., 2013). Next, similarity anchor points were constructed, according to the similarity of sample expression matrixes. Finally, the data were integrated between different studies, according to the identified anchor points.

UGIC Cell Type Identification

After PCA dimensionality reduction was performed on 39,057 UGIC cells, nine cell sets were obtained by T-distributed stochastic neighbor embedding (t-SNE) clustering. In order to identify the cell types, we calculated highly expressed genes on each cluster through the FindMarkers (Butler et al., 2018) function in Seurat. Then, through the artificial gene annotation on the CellMarker (RRID:SCR_018503) database (Zhang X. et al., 2019), the marker genes and the corresponding cell type were finally annotated. We show the statistical graph of cell types identified by EPCAM in the published articles as an example in the CellMarker (RRID:SCR_018503) database (**Supplementary Figure S5**). Then, we analyzed the subtypes of cancer stem cells, obtained a total of six subclasses, and calculated the differentially expressed genes (DEGs) of each subclass.

Other Cancer Cell Type Identification

A total of 71 single-cell sequencing data (**Supplementary Table S2**) from six other cancers were collected. We used the same method to process other tumor single-cell data to ensure the consistency of the analysis process. First, the quality control of single-cell data obtained a total of 215,291 cells. After standardization at the cell level, sample level, and study level, we used PCA and t-SNE visualization to reduce the dimension of those single-cell data and obtained 29 cell collections. We calculated the highly expressed genes of 29 cell collections and used the CellMarker database (Zhang X. et al., 2019) to annotate the cell types. Then, we marked cancer stem cells, which are subtypes 4 and 7. The relevant marker annotations are shown in **Supplementary Figure S3**.

UGIC Transcriptome Sequencing Analysis

We gathered bulk RNA-seq data of UGICs in the TCGA database (Aldape et al., 2015). We obtained the expression matrix data using the cBioPortal (Cerami et al., 2012; Gao et al., 2013), including 522 HNSCC samples, 185 EC samples, and 415 GC samples. Three types of UGICs were congregated with the data label “hnsc_tcga,” “esca_tcga,” and “stad_tcga”. DESeq2 (RRID: SCR_000154) (Love et al., 2014) (1.26.0) software was used to measure the DEGs in the cancer sample and the corresponding normal sample.

Gene Function Annotation

We annotated the function and pathway information of the significantly different genes in the Gene Ontology (GO) (Ashburner et al., 2000; Ashburner, 2021) database and Kyoto Encyclopedia of Genes and Genomes (KEGG) (RRID: SCR_012773) (Kanehisa et al., 2021) database using the clusterProfiler (RRID:SCR_016884) (Yu et al., 2012) (3.14.3) package in R (3.6.2) software. The top 15 terms are presented in **Figure 4**.

Gene Enrichment Analysis

We adjusted the gene set enrichment score between the specific differential genes and the cancer-related gene sets through the Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) software (4.1.0). “C6: oncogenic signatures” was selected as the existing cancer-related gene set in GSEA software. We filtered several parameters to draw gene enrichment results. The normalized enrichment score (NES) was larger than 1. The normalized significance level (NSL) *p*-value was lower than 0.05.

Protein Interaction Network Analysis

We collected all human entries in the String database (RRID: SCR_005223) and deleted low-quality and text mining entries (Szklarczyk et al., 2021). After removing the duplicated edges and self-loops, we constructed a human protein–protein interaction network (PPIN) with 19,267 proteins and 1,689,887 edges by Cytoscape software (RRID:SCR_003032) (Shannon et al., 2003) (3.7.1). Then, 174 genes specifically expressed in UGCSC were mapped to the PPIN. After removing outlier proteins, a regulatory sub-network composed of 144 protein nodes and 545 edges was constructed. Next, we appraised the topological attributes of the network and selected the degree and clustering coefficient (CC) to measure the function of the sub-network (Sporns, 2013). The degree represents the number of connections through a particular node, which measures the importance of the node in the network. CC represents the closeness of connections between a node and the surrounding nodes, which demonstrates the network closeness and function similarity. The **formula (2)** is described as follows:

$$C_i = \frac{2e_i}{d_i\,(d_i - 1)}, \tag{2}$$

where C_i represents the CC of gene i. d_i represents the count of adjacent nodes of gene i. e_i represents the number of interconnected nodes among all adjacent nodes of gene i.

Construction of the Hub Gene Function Network

We manually reviewed the tumor-related literature studies published since 2000 to screen functions and pathways of DEGs. Then, we formulated UGCSC function networks by integrating different genes and the known inflammation and Wnt pathways (**Figure 6**).

RESULT

Landscape of UGIC Single-Cell Data

We collected 39,057 tumor single-cell sequencing data from 30 patients including 4762 HNSCC cells, 366 EC cells, and 33,927 GC cells (**Figure 1A**). It is noteworthy that EC samples applied Smart-seq2 single-cell sequencing technology (Picelli et al., 2014) which is manually sequencing each cell. So the EC group has few cells but higher confidence. After quality filtering (see Methods) and removing the batch effect, more than 70 million transcripts

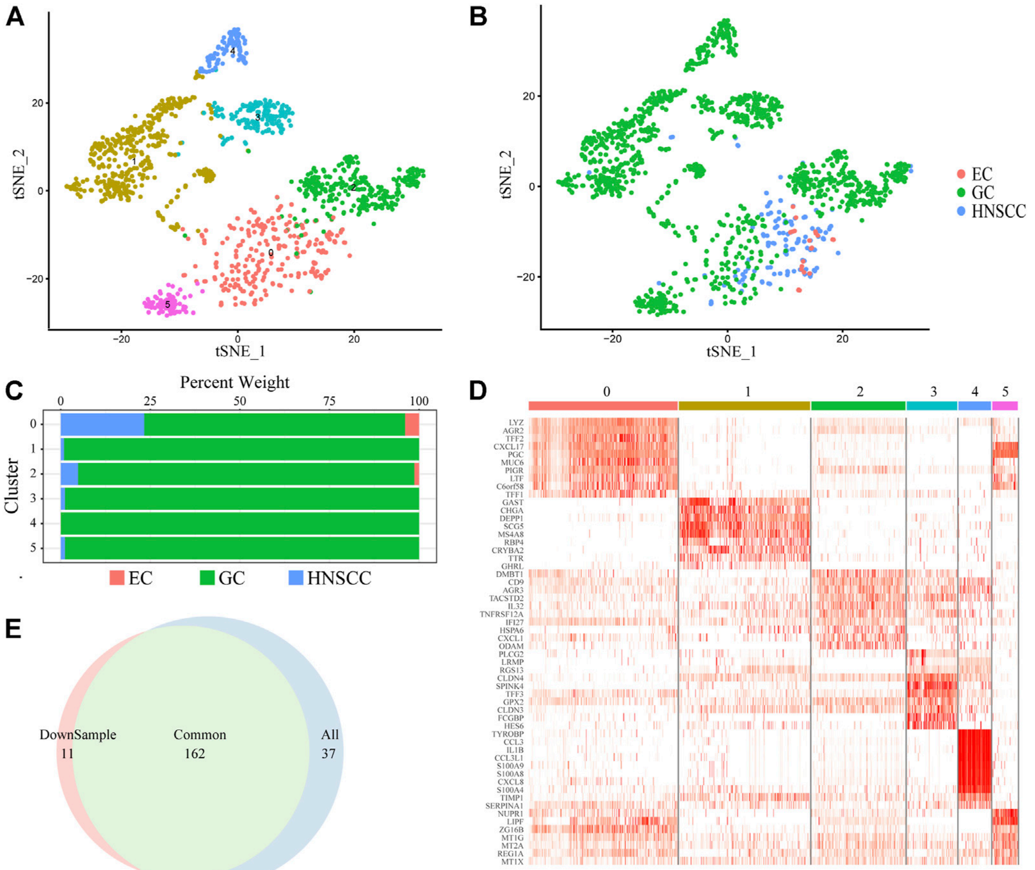


FIGURE 2 | Cancer stem cell clusters. **(A)** t-SNE plot of 1,586 CSC cells color-coded by sub-clusters. **(B)** t-SNE plot color-coded for cancer type of origin. **(C)** Histogram of cell numbers in each CSC cancer type. **(D)** Heatmap of the expression pattern of six sub-clusters in CSC. **(E)** Venn diagram of DEGs in downsample CSCs and All CSCs.

were obtained from 39,057 cells. Subsequently, we classified cells into different clusters by using T-distributed stochastic neighbor embedding (t-SNE) methods in Seurat software (**Supplementary Figure S1**). Through marker genes, these identified cell clusters could be assigned to known cell lineages: T cells, B cells, epithelial cells, natural killer cells, fibroblasts, plasma cells, cancer stem cells, mast cells, and endothelial cells (**Figure 1B**). To corroborate these profiles, we showed the high expression gene distribution heatmap of each cell type and the expression abundance of marker genes of each type (**Figure 1C**; **Supplementary Table S1**). Each cell type has specific marker genes: *CD3D*, *KRT8*, *MS4A1*, *PDGFRA*, *IGHG3*, *EPCAM*, *ECSCR*, *TPSB2*, and *CCL3* (**Figure 1D**). The violin plot of marker genes shows that the expression of most marker genes is specific, which indicates that the classification of cell types is accurate and is very helpful for subsequent analysis (**Figure 1E**). Taken together, these results indicate that the cell classification was accurate, and most of the cells were classified into the correct cell type. The distribution of samples and cancer types is shown in **Supplementary Figures S1, S2**. We also counted the number and frequency of all cell types in HNSCC, EC, and GC and provided the results in **Supplementary Figure S6**.

UGIC-Specific Cancer Stem Cell Identification

We focused on cancer stem cell types in order to reveal the pathogenesis and distant metastasis mechanism of UGIC. We collected a total of 1,586 CSCs (**Figures 2A,B**) including 136 HNSCC cells, 23 EC cells, and 1427 GC cells. Due to the heterogeneity of CSCs, there are differences in the same type of cancer while similarities exist in different types of cancers, coincident with the characteristics of the remote metastasis and recurrence of the cancers. Therefore, we performed a cluster analysis of CSCs, and a total of six sub-clusters were found. After annotating and analyzing all sub-clusters, sub-cluster 0 is ubiquitous in UGICs, including 19 EC stem cells, 356 GC stem cells, and 114 HNSCC stem cells, which proves that sub-cluster 0 preliminarily meets the characteristics of common CSCs (**Figures 2C,D**). Therefore, we concentrated on sub-cluster 0 in the follow-up analysis.

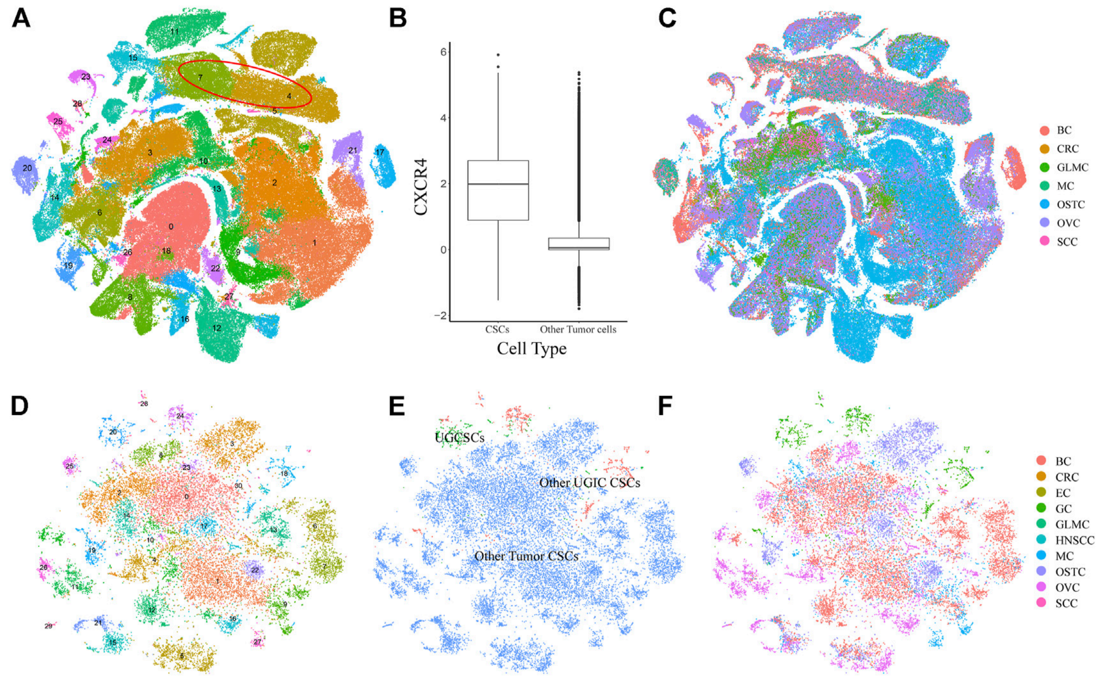


FIGURE 3 | Expression profiling of 215,291 single cells in six cancer types. **(A)** t-SNE plot of 215,291 single cells in six cancer types, color-coded by 29 clusters. Clusters 4 and 7 are cancer stem cells, which are marked by a red circle. **(B)** Box plot shows the expression of the CSC marker gene *CXCR4*. The *x*-axis represents the cell type. The *y*-axis represents the log value of the normalized CXCR4 expression. **(C)** t-SNE plot of other cancer cells, color-coded by the cancer type. **(D)** t-SNE plot of all CSCs, color-coded by clusters. **(E)** t-SNE plot of all CSCs, color-coded by the CSC type. **(F)** t-SNE plot of all CSCs, color-coded by cancer types.

To verify whether sub-cluster 0 reflects the characteristics of UGIC rather than only GC, we performed a down-sampling process in sub-cluster 0 since there are more than 70% GC stem cells in sub-cluster 0. We randomly selected the same number of GC cells as the HNSCC cells and named new sub-cluster 0. Subsequently, we compared the differential genes between sub-cluster 0 and the new sub-cluster 0 in CSCs. The merge ratio is 77.14% (**Figure 2E**), which means these two sub-clusters share the same differential gene set. These results indicate that the differentially expressed genes (DEGs) of sub-cluster 0 represent the features of UGICs.

To validate the specificity of UGCSCs, we compared UGCSCs with other tumor cells. We collected 71 samples (215,291 cells) from six types of cancers including glioma (GLM), melanoma (MELA), osteosarcoma (OSTC), breast cancer (BC), ovarian cancer (OVC), and stellate cell cancer (SCC) (**Supplementary Table S2**). After normalizing the cells and removing the batch effect (see Method), all the cells were gathered into 29 sub-clusters (**Figures 3A,C**). After annotating all cancer cells in the CellMarker database, we noticed that there are plenty of cell types due to the complexity of tissue types involved. Therefore, we only annotated CSCs by using marker genes. The tumor stem cells were obviously aggregated with CXCR4 markers (**Figure 3B**; **Supplementary Figure S3**), which are sub-clusters 4 and 7 and contain 21323 cells, as circled in **Figure 3A**. We compared CSCs of other cancers with CSCs of UGICs. We re-clustered and obtained 31 sub-clusters in all CSCs, which reveals the differences between CSCs of different tumor types (**Figure 3D**). But at the same time, the cluster distribution of CSCs from different tissues is uniform, which indicates that there are similarities between different tissues in CSCs (**Figures 3E,F**). This phenomenon is also coincident with the heterogeneity of tumors. The cluster annotation of cancer types shows that the UGCSC is self-clustering and far away from other tumor CSCs (**Figure 3E**). Therefore, the UGCSC is the specific cancer stem cell in UGIC while UGCSC does not exist in other cancers.

UGCSC Function Analysis

We comprehensively analyzed the distribution and function of UGCSCs. The cell sources of UGCSC cancers were analyzed and counted (**Figure 4A**). As shown in **Figure 4**, UGCSCs are averagely expressed in UGIC patients, including 10 GC patients, four EC patients, and nine HNSCC patients. In summary, the UGCSCs are distributed uniformly, which proves that the UGCSC is common in upper gastrointestinal patients.

We analyzed the expression network of UGCSCs. First, we compared the expression profiles of UGCSCs and all other tumor stem cells and obtained 174 genes with significant differences, including 33 upregulated genes and 141 downregulated genes (**Supplementary Data S1**). We uncovered that the gene information function reflects the characteristics of UGCSCs as a digestive system and as cancer stem cells by analyzing the function annotation (Ashburner et al., 2000) of DEGs (**Figures 4B,D**). The upregulated genes are related to antibacterial response, such as "antibacterial humoral response,"

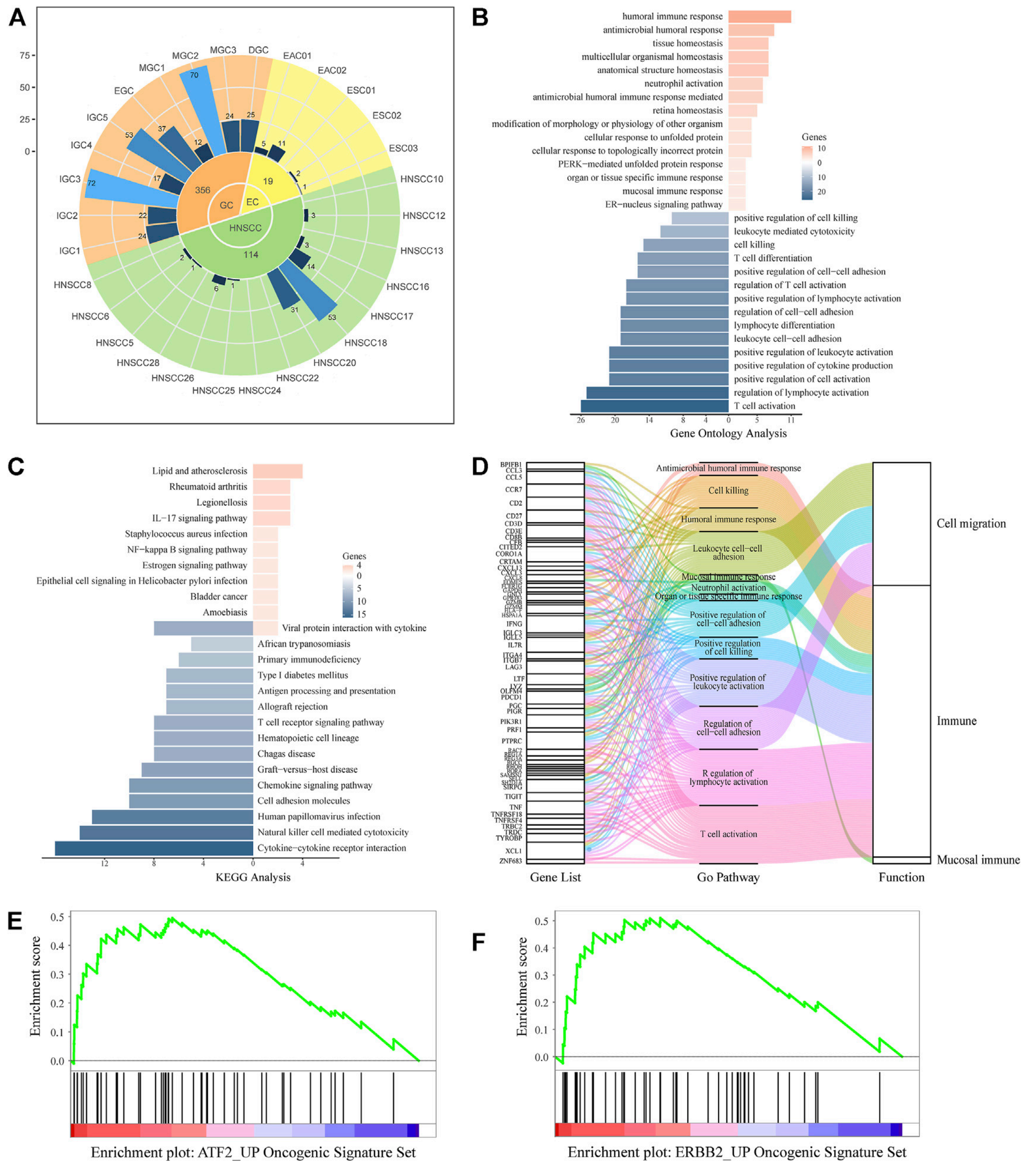


FIGURE 4 | UGCSC function annotation. **(A)** Sample source of UGCSCs. Green, yellow, and orange represent HNSCC, EC, and GC, respectively. The inner circle is for three types of cancer, and the middle circle is the number of cells in each cancer. The outer circle represents the number of cells in each sample. **(B)** Gene ontology function annotation of UGCSC DEGs. **(C)** Pathway analysis of UGCSC differentially expressed genes. **(D)** Gene function integration in UGCSCs. **(E)** UGCSC differentially expressed gene set enrichment analysis in the ATF2_UP tumor set. **(F)** UGCSC differentially expressed gene set enrichment analysis in the ERBB2_UP tumor set.

"antibacterial humoral immune response-mediated," and "mucosal immune response," which are consistent with the function of the digestive tract in the human body. In addition, the functions of downregulated genes mainly focus on reducing the activity of T cells and lymphocytes and downregulating cell killing which could reduce the body's immune response and enhance the survival rate of tumor cells, and these are also the characteristics of cancer stem cells. The downregulated genes also play a role in the regulation of cell–cell adhesion to facilitate the distant metastasis of tumors, which is in line with the feature of metastasis. Through the analysis of the KEGG pathway (Kanehisa et al., 2021), we observed that the significantly differentially expressed genes are enriched in inflammation-related mucosal infections such as "*Staphylococcus aureus* infection," "epithelial cell signaling in *Helicobacter pylori* infection," "IL-17 signaling pathway," and "chemokine signal pathway" (**Figure 4C**). These results uncovered a potential carcinogenic factor of UGIC, that is, mucosal damage induced activation and mutation in

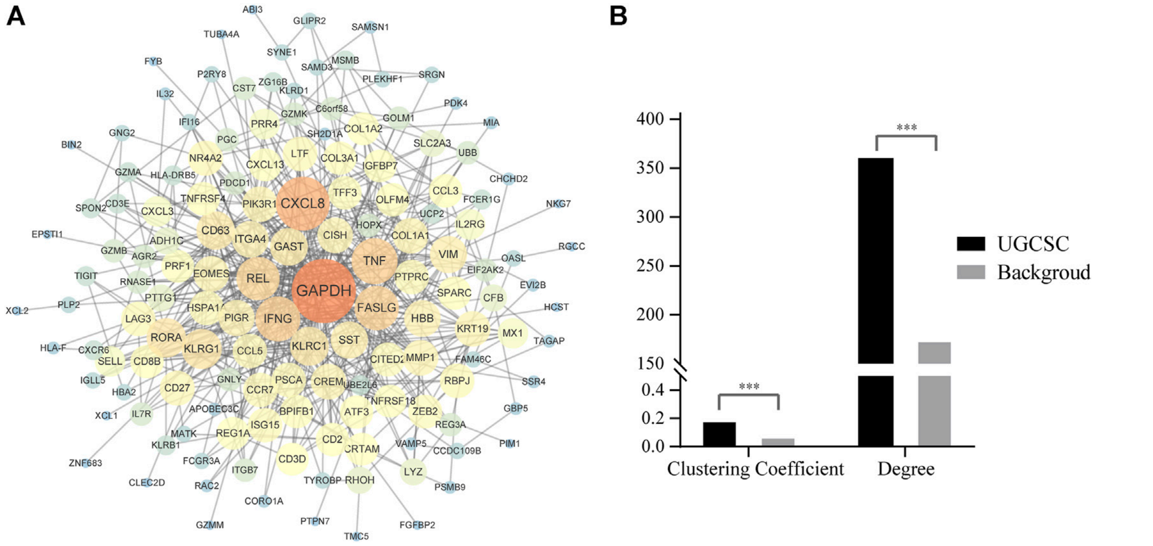


FIGURE 5 | Network analysis of UGCSCs. **(A)** Protein regulation networks of DEGs in UGCSCs. **(B)** Topological attribute comparison of UGCSC sub-networks and whole PPIN.

inflammatory pathways. Through gene set enrichment analysis (GSEA), we found that the significantly differentially expressed genes were significantly enriched in ATF2- and ERBB2-related cancer genes (**Figures 4E,F**). To further confirm the reliability of these genes, we calculated the DEGs between UGCSCs and 30,267 cells of normal tissues in the upper alimentary tract (Cillo et al., 2020; Zhang X. et al., 2021) (**Supplementary Figure S7A**). The DEGs of UGCSCs and tumor cells are few and share fewer genes with the DEGs of UGCSCs and normal cells (**Supplementary Figure S7B**). Also, the functional analysis of DEGs of UGCSCs and normal cells indicates that the functional pathways are related to cell development, which is a common feature of tumors (**Supplementary Figure S7C**). In summary, these results indicate that the DEGs of UGCSCs and tumor cells are oncogenes related to the function of the digestive tract.

UGIC Carcinogenic Mechanism Detection

In order to study the pathogenic mechanisms that may exist in UGCSCs, we mapped 174 proteins into the human protein–protein interaction network (PPIN). We constructed PPIN and deleted low-quality text mining terms in the String database (Szklarczyk et al., 2021). After mapping 174 DEGs in UGCSCs into PPIN, an interaction network consisting of 144 proteins and 545 edges was obtained (**Figure 5A**). Through the analysis of the topological properties of the network, we found that the degree of DEGs in PPIN is 362.174, which is significantly higher than 175.418 in the human total network (**Figure 5B**). This result indicates that the shortest path through different genes is significantly higher than the average value (**Figure 5B**), which implied that these genes are hub genes in the UGCSC network. Furthermore, another topological property, the clustering coefficient (CC), is significantly higher than the background network, which points out that the 144 genes are closely linked compared with the random gene set in the network. The close interaction means a similar or synergistic function in cells. Through the comprehensive analysis of degree and CC, we inferred that the 144 genes are tightly connected hub genes in PPIN, which means that they play an important function in UGCSCs as a co-operative hub gene set.

We have performed functional annotations on the possible functions of these genes and inferred regulatory pathways with the aim to explore the possible pathogenic mechanisms and potential therapeutic targets in UGIC. We analyzed the regulation pathway of those genes through published articles and proved that the upregulated genes are basically related to cancer (**Supplementary Table S3**). Here are some exciting discoveries. Some genes are related to inflammatory pathways, such as CXCL8 (Ha et al., 2017), BPIFB1 (Li J. et al., 2020), PIGR (Kakiuchi et al., 2020), CXCL3, and RNASE1 (Wang et al., 2006), and some genes are related to specific functions of the digestive tract, such as GAST (Giraud et al., 2016), REG1A (Sha et al., 2019), and TFF3 (Braga Emidio et al., 2020). These results illustrated that there may have similar pathogenic mechanisms and common regulatory pathways in some UGICs. We speculated that mucosal damage is induced by long-term unhealthy eating habits, which include smoking, drinking, and hot food breed inflammation. Persistent inflammation leads to carcinogenic mutations and early gastrointestinal tumors. These conjectures have been confirmed in the specific regulatory network of UGCSCs. Based on the detected differentially expressed genes and the mining of relevant research literature studies, we speculated the pathogenesis of the disease, as shown in the **Figure 6**. Inflammation-associated interleukin (CXCL8 and CXCL3) and inflammation defense-related BPIFB1, PIGR, and RNASE1 are activated in UGCSCs. Combined with the epidemiological investigation of gastrointestinal cancer, there is a hypothesis that chronic inflammation is incited by mucosal damage due to long-term bad eating habits. We present that the cancerous chronic inflammation is activated by GAST, REG1A, TFF3, and ZG16B in the Wnt signaling pathway. Upregulated hPG80 and TFF3

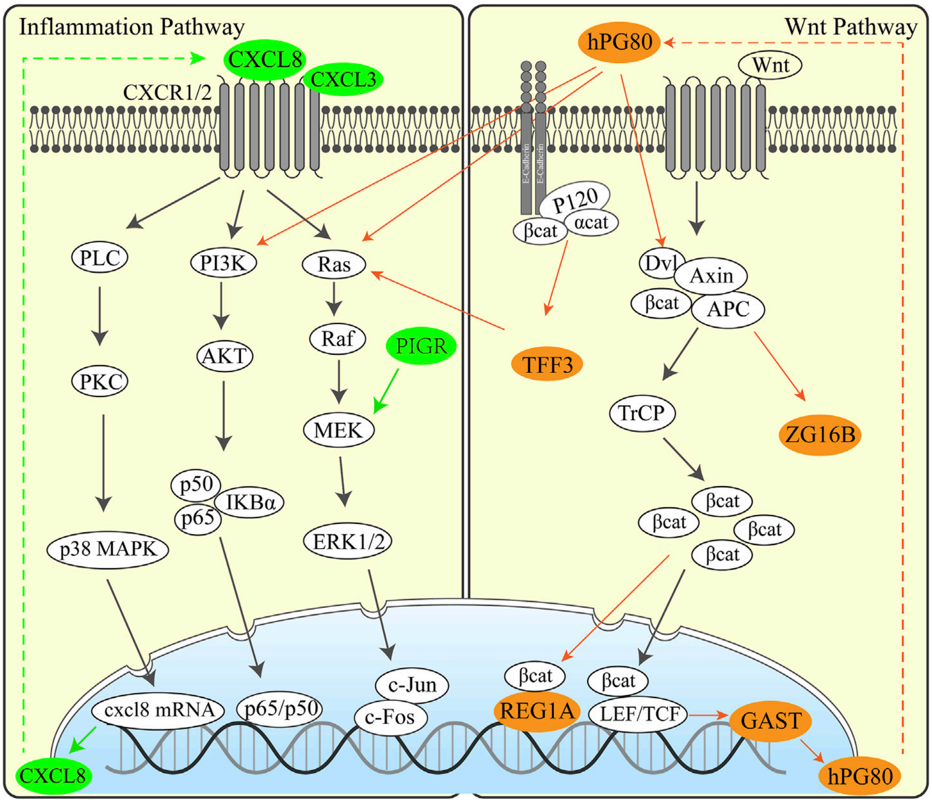


FIGURE 6 | Regulation pathway in UGCSCs. Inflammatory pathway (left) and Wnt pathway (right) activated in UGCSC. The green genes represent inflammatory factors that are highly expressed in UGCSCs. The brown–yellow genes represent Wnt-related factors that are highly expressed in UGCSCs.

induce PI3K/Ras and lead to tumor cell growth and invasion, which may be one reason for the poor prognosis of UGIC. Hence, this resource provides a novel view for the occurrence and development of UGIC and the advancement of gastrointestinal cancer diagnosis and therapy.

CONCLUSION

In this article, we collectively analyzed single-cell sequencing data of HNSCC, EC, and GC and identified a specific cancer stem cell type in UGIC: UGCSCs. Then, we presented the unique expression pattern and hub gene set in UGCSCs by comparing it with other tumors' single-cell RNA-seq data. We declared the common carcinogens of UGICs that the mucosa damage of the digestive tract induces chronic inflammation due to unhealthy eating habits. The hub gene set provides promising entry points for the design of novel therapies including CXCL8, CXCL3, GAST, TFF3, PIGR, and RNASE1.

DISCUSSION

Here, we provided a comprehensive catalog of human UGICs at single-cell resolution. In the integrative analysis of UGICs, we confirmed that there are specific cancer stem cells in UGIC, which are named UGCSCs. This discovery provides a new perspective for scientific analysis of the poor prognosis and easy recurrence of UGIC. By comparing the tumor stem cells of six cancers, we extracted the core gene set that plays an important role in UGCSCs and explored the possible pathogenic pathway of UGIC and core genes including *GAST*, *CXCL8*, *CXCL3*, *PIGR*, *REG1A*, and *TFF3*. With further in-depth research, these genes can also be used as diagnostic markers or possible therapeutic targets for gastrointestinal cancers.

However, all cell types and subtypes cannot possibly be described here; some key results emerge. On one hand, the distribution of all cell types in UGIC is shown in the cell clustering figure (**Figure 1B**). On the other hand, through the comparative analysis with bulk RNA-seq sequencing data, the DEGs between single-cell data and bulk RNA-seq data varied significantly. Therefore, we performed further research only on cancer stem cells. Intriguing questions remain as to whether there are specific immune cells in UGIC and whether the immune cell counts would have an impact on the prognosis of UGIC.

The single-cell data of UGIC and the six cancer types are composed of cells from different patients. Some sub-clusters of cell types have different abundances due to sample differences, according to the results of cell clustering. We removed batch effects and deleted outlier cells from the clustering result. In this way, the impact of samples from different patient sources is reduced.

We performed the same analysis on bulk RNA sequencing data; however, due to the varieties of cell types and the low proportion of CSCs in cancer tissues, the pathways and therapeutic targets were not discovered. We collected 1,122 patients and 1,966 normal samples of bulk RNA-seq data in TCGA database. The differentially expressed genes of the three cancers were compared with those of UGCSCs. The result suggests that the merge ratio is only 0.79%. Moreover, the function of differentially expressed genes is mostly about cell cycle-related pathways in bulk RNA-seq data (**Supplementary Figure S4**). We inferred that plenty of cell types in UGIC generates noises in UGIC expression profile information and makes some core pathways and genes undetectable, while single-cell RNA-seq can filter noise signals by extracting specific cell types.

Last, we constructed a regulatory network of UGCSCs under the framework of the existing experimental knowledge atlas. More and other types of data such as downstream genes and mutation information of the core regulatory network need to be further studied. However, we proposed UGCSCs and their regulatory networks based on the analysis of single-cell data from more than 100 patients and more than 25,000 cells, which has strong robustness. These data build a framework for a deeper understanding of the molecular mechanisms of UGCSCs and the regulation network of hub genes and might be applied to screen for molecular target drugs to improve the efficacy and outcomes for UGIC patients.

DATA AVAILABILITY STATEMENT

The UGIC single-cell sequencing data are available through SRR6133148 (Wu et al., 2018), GSE103322 (Puram et al., 2017), and GSE134520 (Zhang et al., 2019a) in the Gene Expression Omnibus (GEO) and HRA000051 (Zhang et al., 2021a) in Genome Sequence Archive (GSA). The normal single cell data of upper alimentary tract are available through GSE139324 (Cillo et al., 2020) and GSE160269 (Zhang et al., 2021b) in the GEO database and HRA000051 (Zhang et al., 2021a) in GSA database. The other cancer single-cell sequencing data are available through GEO database: GSE152048 (Zhou et al., 2020), GSE89567 (Venteicher et al., 2017), GSE72056 (Tirosh et al., 2016), GSE75688 (Chung et al., 2017), GSE84465 (Darmanis et al., 2017), and GSE102130 (Filbin et al., 2018) in GEO database and https://lambrechtslab.sites.vib.be/ (Qian et al., 2020) website. Gene interactions networks were identified using the STRING database (https://string-db.org) (Szklarczyk et al., 2021). GO term analysis was performed by using the GENEONTOLOGY database (http://geneontology.org/) (Ashburner et al., 2000). Tumor gene sets enrichment analysis was performed using the MSigDB database (www.gseamsigdb.org) (Subramanian et al., 2005). All other data are included in the article and its **Supplementary Information** files or available from the corresponding authors upon reasonable request.

AUTHOR CONTRIBUTIONS

Study concept and design: LL and HL; acquisition of data: LL and YjZ; analysis and interpretation of data: LL, YjZ, YR, ZC, YnZ, XW, HZ, and HL; drafting of the manuscript: LL, HZ, and HL. All authors reviewed and approved the final draft of the manuscript.

FUNDING

This work was supported by the National Key R&D Program of China (2018YFC0910500), the SJTU-Yale Collaborative Research Seed Fund, and the Neil Shen's SJTU Medical Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.849798/full#supplementary-material

REFERENCES

Akhtar, S. (2013). Areca Nut Chewing and Esophageal Squamous-Cell Carcinoma Risk in Asians: A Meta-Analysis of Case-Control Studies. *Cancer Causes Control* 24, 257–265. doi:10.1007/s10552-012-0113-9

Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G., and von Deimling, A. (2015). Glioblastoma: Pathology, Molecular Mechanisms and Markers. *Acta Neuropathol.* 129, 829–848. doi:10.1007/s00401-015-1432-1

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). "Deep Canonical Correlation Analysis," in International Conference on Machine Learning, PMLR, Atlanta, GA, USA, Jun 16-21, 2013, 1247–1255.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Ashburner (2021). The Gene Ontology Resource: Enriching a GOld Mine. *J. Nucl. Acids Res.* 49 (D1), D325–D334.

Babina, I. S., and Turner, N. C. (2017). Advances and Challenges in Targeting FGFR Signalling in Cancer. *Nat. Rev. Cancer* 17, 318–332. doi:10.1038/nrc.2017.8

Braga Emidio, N., Brierley, S. M., Schroeder, C. I., and Muttenthaler, M. (2020). Structure, Function, and Therapeutic Potential of the Trefoil Factor Family in the Gastrointestinal Tract. *ACS Pharmacol. Transl. Sci.* 3, 583–597. doi:10.1021/acsptsci.0c00023

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* 68, 394–424. doi:10.3322/caac.21492

Bu, X., Juneja, V. R., Reynolds, C. G., Mahoney, K. M., Bu, M. T., McGuire, K. A., et al. (2021). Monitoring PD-1 Phosphorylation to Evaluate PD-1 Signaling during Antitumor Immune Responses. *Cancer Immunol. Res.* 9, 1465–1475. doi:10.1158/2326-6066.CIR-21-0493

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species. *Nat. Biotechnol.* 36, 411–420. doi:10.1038/nbt.4096

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* 2, 401–404. doi:10.1158/2159-8290.CD-12-0095

Chakravarthy, A., Furness, A., Joshi, K., Ghorani, E., Ford, K., Ward, M. J., et al. (2018). Pan-cancer Deconvolution of Tumour Composition Using DNA Methylation. *Nat. Commun.* 9, 1–13. doi:10.1038/s41467-018-05570-1

Chung, W., Eum, H. H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., et al. (2017). Single-cell RNA-Seq Enables Comprehensive Tumour and Immune Cell Profiling in Primary Breast Cancer. *Nat. Commun.* 8, 1–12. doi:10.1038/ncomms15081

Cillo, A. R., Kürten, C. H. L., Tabib, T., Qi, Z., Onkar, S., Wang, T., et al. (2020). Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity* 52, 183–199. doi:10.1016/j.immuni.2019.11.014

Clarke, M. F., Dick, J. E., Dirks, P. B., Eaves, C. J., Jamieson, C. H. M., Jones, D. L., et al. (2006). Cancer Stem Cells-Perspectives on Current Status and Future Directions: AACR Workshop on Cancer Stem Cells. *Cancer Res.* 66, 9339–9344. doi:10.1158/0008-5472.CAN-06-3126

Cui, Y., Guo, W., Li, Y., Shi, J., Ma, S., and Guan, F. (2021). Pan-cancer Analysis Identifies ESM1 as a Novel Oncogene for Esophageal Cancer. *Esophagus* 18, 326–338. doi:10.1007/s10388-020-00796-9

Dai, L., Huang, Z., and Li, W. (2021). Analysis of the PD-1 Ligands Among Gastrointestinal Cancer Patients: Focus on Cancer Immunity. *Front. Oncol.* 11, 525. doi:10.3389/fonc.2021.637015

Darmanis, S., Sloan, S. A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., et al. (2017). Single-cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cel Rep.* 21, 1399–1410. doi:10.1016/j.celrep.2017.10.030

Dhakras, P., Uboha, N., Horner, V., Reinig, E., and Matkowskyj, K. A. (2020). Gastrointestinal Cancers: Current Biomarkers in Esophageal and Gastric Adenocarcinoma. *Transl. Gastroenterol. Hepatol.* 5, 55. doi:10.21037/tgh.2020.01.08

Filbin, M. G., Tirosh, I., Hovestadt, V., Shaw, M. L., Escalante, L. E., Mathewson, N. D., et al. (2018). Developmental and Oncogenic Programs in H3K27M Gliomas Dissected by Single-Cell RNA-Seq. *Science* 360, 331–335. doi:10.1126/science.aao4750

Freytag, S., Tian, L., Lönnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of Clustering Tools in R for Medium-Sized 10x Genomics Single-Cell RNA-Sequencing Data. *F1000Res* 7, 1297. doi:10.12688/f1000research.15809.2

Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P., et al. (2008). Tobacco Smoking and Cancer: A Meta-Analysis. *Int. J. Cancer* 122, 155–164. doi:10.1002/ijc.23033

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* 6, pl1. doi:10.1126/scisignal.2004088

Giraud, J., Failla, L. M., Pascussi, J.-M., Lagerqvist, E. L., Ollier, J., Finetti, P., et al. (2016). Autocrine Secretion of Progastrin Promotes the Survival and Self-Renewal of Colon Cancer Stem-like Cells. *Cancer Res.* 76, 3618–3628. doi:10.1158/0008-5472.CAN-15-1497

Goldstein, B. Y., Chang, S.-C., Hashibe, M., La Vecchia, C., and Zhang, Z.-F. (2010). Alcohol Consumption and Cancers of the Oral Cavity and Pharynx from 1988 to 2009: an Update. *Eur. J. Cancer Prev.* 19, 431–465. doi:10.1097/CEJ.0b013e32833d936d

González, C. A., Sala, N., and Rokkas, T. (2013). Gastric Cancer: Epidemiologic Aspects. *Helicobacter* 18, 34–38. doi:10.1111/hel.12082

Ha, H., Debnath, B., and Neamati, N. (2017). Role of the CXCL8-CXCR1/2 axis in Cancer and Inflammatory Diseases. *Theranostics* 7, 1543–1588. doi:10.7150/thno.15625

Haas, S. L., Ye, W., and Löhr, J.-M. (2012). Alcohol Consumption and Digestive Tract Cancer. *Curr. Opin. Clin. Nutr. Metab. Care* 15, 457–467. doi:10.1097/MCO.0b013e3283566699

Johnson, D. E., O'Keefe, R. A., and Grandis, J. R. (2018). Targeting the IL-6/JAK/STAT3 Signalling axis in Cancer. *Nat. Rev. Clin. Oncol.* 15, 234–248. doi:10.1038/nrclinonc.2018.8

Kakiuchi, N., Yoshida, K., Uchino, M., Kihara, T., Akaki, K., Inoue, Y., et al. (2020). Frequent Mutations that Converge on the NFKBIZ Pathway in Ulcerative Colitis. *Nature* 577, 260–265. doi:10.1038/s41586-019-1856-1

Kanehisa, M., Sato, Y., and Kawashima, M. (2021). KEGG Mapping Tools for Uncovering Hidden Features in Biological Data. *Protein Sci.* 31, 47–53. doi:10.1002/pro.4172

Karakasheva, T. A., Lin, E. W., Tang, Q., Qiao, E., Waldron, T. J., Soni, M., et al. (2018). IL-6 Mediates Cross-Talk between Tumor Cells and Activated Fibroblasts in the Tumor Microenvironment. *Cancer Res.* 78, 4957–4970. doi:10.1158/0008-5472.CAN-17-2268

Li, J., Xu, P., Wang, L., Feng, M., Chen, D., Yu, X., et al. (2020a). Molecular Biology of BPIFB1 and its Advances in Disease. *Ann. Transl. Med.* 8, 651. doi:10.21037/atm-20-3462

Li, Z., Chen, S., Feng, W., Luo, Y., Lai, H., Li, Q., et al. (2020b). A Pan-Cancer Analysis of HER2 index Revealed Transcriptional Pattern for Precise Selection of HER2-Targeted Therapy. *EBioMedicine* 62, 103074. doi:10.1016/j.ebiom.2020.103074

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15, 1–21. doi:10.1186/s13059-014-0550-8

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-Seq from Single Cells Using Smart-Seq2. *Nat. Protoc.* 9, 171–181. doi:10.1038/nprot.2014.006

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* 171, 1611–1624. doi:10.1016/j.cell.2017.10.044

Qian, J., Olbrecht, S., Boeckx, B., Vos, H., Laoui, D., Etlioglu, E., et al. (2020). A Pan-Cancer Blueprint of the Heterogeneous Tumor Microenvironment Revealed by Single-Cell Profiling. *Cell Res.* 30, 745–762. doi:10.1038/s41422-020-0355-0

Rahman, S. A., Yagnik, B., Bally, A. P., Morrow, K. N., Wang, S., Vanderford, T. H., et al. (2021). PD-1 Blockade and Vaccination Provide Therapeutic Benefit against SIV by Inducing Broad and Functional CD8 + T Cells in Lymphoid Tissue. *Sci. Immunol.* 6, eabh3034. doi:10.1126/sciimmunol.abh3034

Saleh, R. R., Peinado, P., Fuentes-Antrás, J., Pérez-Segura, P., Pandiella, A., Amir, E., et al. (2019). Prognostic Value of Lymphocyte-Activation Gene 3 (LAG3) in Cancer: a Meta-Analysis. *Front. Oncol.* 9, 1040. doi:10.3389/fonc.2019.01040

Sha, Y.-L., Liu, S., Yan, W.-W., and Dong, B. (2019). Wnt/β-catenin Signaling as a Useful Therapeutic Target in Hepatoblastoma. *Biosci. Rep.* 39, BSR20192466. doi:10.1042/BSR20192466

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

Sporns, O. (2013). Network Attributes for Segregation and Integration in the Human Brain. *Curr. Opin. Neurobiol.* 23, 162–171. doi:10.1016/j.conb.2012.11.015

Sreepadmanabh, M., and Toley, B. J. (2018). Investigations into the Cancer Stem Cell Niche Using Iin-Vvitro 3-D Tumor Models and Microfluidics. *Biotechnol. Adv.* 36, 1094–1110. doi:10.1016/j.biotechadv.2018.03.009

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., III, et al. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888–1902. doi:10.1016/j.cell.2019.05.031

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/nar/gkaa1074

Thienpont, B., Steinbacher, J., Zhao, H., D'Anna, F., Kuchnio, A., Ploumakis, A., et al. (2016). Tumour Hypoxia Causes DNA Hypermethylation by Reducing TET Activity. *Nature* 537, 63–68. doi:10.1038/nature19081

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq. *Science* 352, 189–196. doi:10.1126/science.aad0501

Tran, E., Ahmadzadeh, M., Lu, Y.-C., Gros, A., Turcotte, S., Robbins, P. F., et al. (2015). Immunogenicity of Somatic Mutations in Human Gastrointestinal Cancers. *Science* 350, 1387–1390. doi:10.1126/science.aad1253

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., et al. (2017). Decoupling Genetics, Lineages, and Microenvironment in IDH-Mutant Gliomas by Single-Cell RNA-Seq. *Science* 355, eaai8478. doi:10.1126/science.aai8478

Wang, L., Zhu, J.-S., Song, M.-Q., Chen, G.-Q., and Chen, J.-L. (2006). Comparison of Gene Expression Profiles between Primary Tumor and Metastatic Lesions in Gastric Cancer Patients Using Laser Microdissection and cDNA Microarray. *World J. Gastroenterol.* 12, 6949. doi:10.3748/wjg.v12.i43.6949

Wang, Z., Yin, N., Zhang, Z., Zhang, Y., Zhang, G., and Chen, W. (2017). Upregulation of T-Cell Immunoglobulin and Mucin-Domain Containing-3 (Tim-3) in Monocytes/macrophages Associates with Gastric Cancer Progression. *Immunol. Invest.* 46, 134–148. doi:10.1080/08820139.2016.1229790

Westphalen, C. B., Preinfalk, A., Kruger, S., Haas, M., Renz, B. W., Riener, M.-O., et al. (2019). Neurotrophic Tropomyosin Receptor Kinase (NTRK) and Nerve Growth Factor (NGF) Are Not Expressed in Caucasian Patients with Biliary Tract Cancers: Pooled Data from Three Independent Cohorts. *Clin. Transl. Oncol.* 21, 1108–1111. doi:10.1007/s12094-018-02030-6

Wu, H., Yu, J., Li, Y., Hou, Q., Zhou, R., Zhang, N., et al. (2018). Single-cell RNA Sequencing Reveals Diverse Intratumoral Heterogeneities and Gene Signatures of Two Types of Esophageal Cancers. *Cancer Lett.* 438, 133–143. doi:10.1016/j.canlet.2018.09.017

Yamada, T., Alpers, D. H., Kalloo, A. N., Kaplowitz, N., Owyang, C., and Powell, D. W. (2011). *Textbook of Gastroenterology*. Seattle, Washington: John Wiley & Sons.

Yang, Y., Zhang, J., Chen, Y., Xu, R., Zhao, Q., and Guo, W. (2020). MUC4, MUC16, and TTN Genes Mutation Correlated with Prognosis, and Predicted Tumor Mutation burden and Immunotherapy Efficacy in Gastric Cancer and pan-cancer. *Clin. Transl. Med.* 10, e155. doi:10.1002/ctm2.155

Yin, J., Wu, X., Li, S., Li, C., and Guo, Z. (2020). Impact of Environmental Factors on Gastric Cancer: A Review of the Scientific Evidence, Human Prevention and Adaptation. *J. Environ. Sci.* 89, 65–79. doi:10.1016/j.jes.2019.09.025

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* 16, 284–287. doi:10.1089/omi.2011.0118

Zhang, H.-Z., Jin, G.-F., and Shen, H.-B. (2012). Epidemiologic Differences in Esophageal Cancer between Asian and Western Populations. *Chin. J. Cancer* 31, 281–286. doi:10.5732/cjc.011.10390

Zhang, P., Yang, M., Zhang, Y., Xiao, S., Lai, X., Tan, A., et al. (2019a). Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. *Cel Rep.* 27, 1934e1935–1947. doi:10.1016/j.celrep.2019.04.052

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., et al. (2019b). CellMarker: a Manually Curated Resource of Cell Markers in Human and Mouse. *Nucleic Acids Res.* 47, D721–D728. doi:10.1093/nar/gky900

Zhang, J., Dong, R., Dong, R., and Shen, L. (2020). Evaluation and Reflection on Claudin 18.2 Targeting Therapy in Advanced Gastric Cancer. *Chin. J. Cancer Res.* 32, 263–270. doi:10.21147/j.issn.1000-9604.2020.02.13

Zhang, M., Hu, S., Min, M., Ni, Y., Lu, Z., Sun, X., et al. (2021a). Dissecting Transcriptional Heterogeneity in Primary Gastric Adenocarcinoma by Single Cell RNA Sequencing. *Gut* 70, 464–475. doi:10.1136/gutjnl-2019-320368

Zhang, X., Peng, L., Luo, Y., Zhang, S., Pu, Y., Chen, Y., et al. (2021b). Dissecting Esophageal Squamous-Cell Carcinoma Ecosystem by Single-Cell Transcriptomic Analysis. *Nat. Commun.* 12, 1–17. doi:10.1038/s41467-021-25539-x

Zhang, Y. (2013). Epidemiology of Esophageal Cancer. *World J. Gastroenterol.* 19, 5598. doi:10.3748/wjg.v19.i34.5598

Zhou, Y., Yang, D., Yang, Q., Lv, X., Huang, W., Zhou, Z., et al. (2020). Single-cell RNA Landscape of Intratumoral Heterogeneity and Immunosuppressive Microenvironment in Advanced Osteosarcoma. *Nat. Commun.* 11, 1–17. doi:10.1038/s41467-020-20059-6

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 | Frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 30 May 2022
doi: 10.3389/fbioe.2022.916309

Identifying In Vitro Cultured Human Hepatocytes Markers with Machine Learning Methods Based on Single-Cell RNA-Seq Data

ZhanDong Li[1†], *FeiMing Huang*[2†], *Lei Chen*[3†], *Tao Huang*[4,5]* *and Yu-Dong Cai*[2]*

[1]*College of Biological and Food Engineering, Jilin Engineering Normal University, Changchun, China,* [2]*School of Life Sciences, Shanghai University, Shanghai, China,* [3]*College of Information Engineering, Shanghai Maritime University, Shanghai, China,* [4]*Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China,* [5]*CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China*

Cell transplantation is an effective method for compensating for the loss of liver function and improve patient survival. However, given that hepatocytes cultivated *in vitro* have diverse developmental processes and physiological features, obtaining hepatocytes that can properly function *in vivo* is difficult. In the present study, we present an advanced computational analysis on single-cell transcriptional profiling to resolve the heterogeneity of the hepatocyte differentiation process *in vitro* and to mine biomarkers at different periods of differentiation. We obtained a batch of compressed and effective classification features with the Boruta method and ranked them using the Max-Relevance and Min-Redundancy method. Some key genes were identified during the *in vitro* culture of hepatocytes, including *CD147*, which not only regulates terminally differentiated cells in the liver but also affects cell differentiation. *PPIA*, which encodes a CD147 ligand, also appeared in the identified gene list, and the combination of the two proteins mediated multiple biological pathways. Other genes, such as *TMSB10*, *TMEM176B*, and *CD63*, which are involved in the maturation and differentiation of hepatocytes and assist different hepatic cell types in performing their roles were also identified. Then, several classifiers were trained and evaluated to obtain optimal classifiers and optimal feature subsets, using three classification algorithms (random forest, k-nearest neighbor, and decision tree) and the incremental feature selection method. The best random forest classifier with a 0.940 Matthews correlation coefficient was constructed to distinguish different hepatic cell types. Finally, classification rules were created for quantitatively describing hepatic cell types. In summary, This study provided potential targets for cell transplantation associated liver disease treatment strategies by elucidating the process and mechanism of hepatocyte development at both qualitative and quantitative levels.

Keywords: hepatocytes, single cell RNA sequencing, machine learning, boruta, max-relevance, min-redundancy and random forest

OPEN ACCESS

Edited by:
Jiaofang Shao,
Nanjing Medical University, China

Reviewed by:
Jin Deng,
South China Agricultural University, China
Jing Yang,
ShanghaiTech University, China
Azian Azamimi Abdullah,
Universiti Malaysia Perlis, Malaysia

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *09 April 2022*
Accepted: *11 May 2022*
Published: *30 May 2022*

Citation:
Li Z, Huang F, Chen L, Huang T and Cai Y-D (2022) Identifying In Vitro Cultured Human Hepatocytes Markers with Machine Learning Methods Based on Single-Cell RNA-Seq Data. Front. Bioeng. Biotechnol. 10:916309. doi: 10.3389/fbioe.2022.916309

INTRODUCTION

Over the past few decades, liver disease has gradually become one of the leading causes of death worldwide. Acute hepatitis, cirrhosis, and liver cancer account for approximately 4% of all deaths globally (Xiao et al., 2019). The only treatment for an end-stage liver disease that impairs the ability of the liver to regenerate is liver transplantation (Zhang et al., 2018). However, the practical use of liver transplantation is limited by the shortage of liver grafts for transplantation (Iansante et al., 2018). A potential alternative therapy for liver transplantation, allogeneic hepatocyte transplantation requires the cultivation of active hepatocytes *in vitro* (Iansante et al., 2018). However, obtaining hepatocytes that can function properly *in vivo* is difficult because of the different developmental processes and physiological characteristics of hepatocytes cultured *in vitro* (Hu and Li, 2015). Therefore, the development of functional hepatocytes for liver regeneration is a priority. The developmental mechanisms and heterogeneous characteristics of hepatocytes *in vitro* have become major subjects of interest because of the high clinical demand.

Liver transplant patients experience alloimmune rejection, which may cause various complications and affect the long-term survival of recipients (Du et al., 2020). Chronic allograft injury, late graft failure, and the negative effects of anti-rejection medication continue to be the major roadblocks to good outcomes (Thomson et al., 2020). Following the development of allogeneic hepatocyte transplantation technology, analysis methods for hepatic cell types and immune cell characteristics *in vitro* have become effective tools for the study of immune rejection (Kawahara et al., 1998; Iansante et al., 2018). Different hepatic cell types, including hepatoblasts, hepatocytes, and cholangiocytes, which are cultured *in vitro* and can be transplanted into a damaged liver, can repair the liver and improve liver function. The challenge of culturing functional hepatocytes *in vitro* is enormous. Primary hepatocytes have difficulty maintaining stimulation by a complex set of factors *in vivo* during *in vitro* culture, resulting in loss of hepatocyte polarity and function (Lauschke et al., 2019). In addition, owing to the shortage of donors and the lack of strategies that can increase these donors, primary hepatocytes are extremely scarce to meet the conditions for treatment. The selection of appropriate original stem cells and an *in vitro* system suitable for stem cell differentiation is crucial to the differentiation of stem cells into mature liver type cells (Guo et al., 2017). It is particularly significant to explore the process of differentiation of different original stem cells *in vitro* and to elucidate the key pathways that maintain the properties of primary hepatocytes.

Through single-cell sequencing, scientists can now investigate the mechanisms of cell growth and differentiation in unprecedented detail and resolve cell heterogeneity. Aizarani et al. successfully resolved the heterogeneity of human hepatocytes *in vivo* and the differentiation process (Aizarani et al., 2019). However, owing to environmental differences, hepatocytes cultured *in vitro* can show characteristics different from those cultured *in vivo*. Logan et al. distinguish hepatocytes cultured *in vitro* on the basis of cell shape with a machine learning approach (Logan et al., 2016). However, distinguishing hepatocytes at different stages of differentiation *in vitro* by this method remains difficult because of the diversity and ambiguity of cell morphology during development. In our study, the transcriptional profiles of different hepatic cell types cultured *in vitro* are combined using advanced machine learning methods, and the characteristic markers of various hepatocyte populations were identified. Results suggest the functional characteristics of each population. Advanced computational methods for describing liver cells cultured *in vitro* and resolving hepatocyte developmental processes and mechanisms have become a focus of research as the amount and variety of data grow.

Here, we uncovered a series of genes and classification rules linked with *in vitro* hepatocyte differentiation processes and type specificity by using advanced computational approaches based on public single-cell RNA sequencing data. First, we used two effective feature selection approaches (Boruta (Kursa and Rudnicki, 2010) and Max-Relevance and Min-Redundancy (Peng et al., 2005)) to filter and rank features. Based on ranked features, several feature sets were constructed in incremental feature selection (IFS) approaches (Liu and Setiono, 1998), which were fed into three efficient classification algorithms to build classifiers. The optimal classifier and the optimal feature subset were obtained by evaluating the performance of each classifier and observing the IFS curve. A number of genes in the optimal feature subset are associated with hepatocyte differentiation and function, demonstrating the accuracy of our computational analysis. In addition, a series of quantitative rules were established for distinguishing specific cell types and functions during hepatocyte differentiation *in vitro*. Overall, our study provided a novel computational analysis for revealing the characteristic markers of various hepatocyte populations, suggesting the functional characteristics of each cell population. The top-ranked features and decision rules identified by our analysis provided a theoretical basis for resolving hepatocyte developmental processes and mechanisms and potential targets for the treatment of clinical liver diseases.

MATERIALS AND METHODS

Data

We obtained *in vitro* cultured human hepatocyte single-cell RNA sequencing expression profiles from the Gene Expression Omnibus (GEO) database under accession number GSE128060 (Feng S. et al., 2020). These data include 1,147 cells from 16 different hepatic cell types, each with 63,255 genes at different expression levels obtained through Smart-Seq2 sequencing. The sample sizes of each hepatic cell type are listed in **Table 1**. In each cell, the expression levels of genes were quantified using the transcript-per-million method.

Boruta Feature Filtering

The majority of the features is irrelevant to the classification. When all features are selected for further analysis, redundancy and noise are introduced, which might lead to biased calculations.

TABLE 1 | The sample sizes of different cell types cultured *in vitro*.

Class Index	Cell types	Sample size
1	5C-condition cultured human primary hepatocyte	96
2	Cultured human primary intrahepatic biliary epithelial cell	34
3	Definitive endoderm	15
4	Endoderm stem cell (EnSC)	24
5	EnSC-derived cholangiocyte	68
6	EnSC-derived EGFi-untreated hepatocyte	128
7	EnSC-derived hepatic endoderm	59
8	EnSC-derived hepatoblast	84
9	EnSC-derived hepatocyte	177
10	EnSC-derived immature hepatocyte	31
11	EnSC-derived TPPB-untreated cholangiocyte	75
12	Hepatocyte derived from ProliHH P2 through 3D maturation	22
13	Hepatocyte derived from ProliHH P5 through 3D maturation	32
14	Human embryonic stem cell-derived hepatocyte-like cell	140
15	Sorted ALB+ CYP3A4+ EnSC-derived hepatocyte	67
16	Uncultured adult human primary hepatocyte	95

We used the Boruta approach to filter extraneous features in this case (Kursa and Rudnicki, 2010). The Boruta feature filtering method has been widely used in biological data mining in the past (Chen L. et al., 2021; Ding et al., 2021).

Boruta is based on the random forest (RF) classifier, which adds randomness to a system and collects results from a collection of random features. This function reduces the misleading effects of random fluctuations and correlations for the generation of the most relevant features for classification. Boruta includes the following steps: *1*) When modeling for the first time, copies of the original variables as shadow variables are generated. *2*) The values of the corresponding shadow variables are randomly shuffled. *3*) The importance score of each variable is calculated with RF modeling. *4*) For each true characteristic variable, the difference between its significance maximum and that of each shadow variable is evaluated using statistical tests. The true characteristic variables with significantly higher importance than the shadow variables are defined as significant. Real characteristic variables with significantly lower importance than the shadow variables are defined as insignificant. *5*) All insignificant variables and shadow variables are removed. The modeling and selection process is repeated and performed on the basis of the new variable composition of the dataset until all variables are classified as significant or insignificant, or a pre-set number of iterations is reached.

We used the Boruta tool from https://github.com/scikit-learn-contrib/boruta_py in this study and used the default parameters for the analysis.

Max-Relevance and Min-Redundancy

mRMR is a filtered feature selection algorithm that maximizes the relevance between features and targets and decreases the redundancy between selected features (Peng et al., 2005; Zhu et al., 2020; Chen et al., 2022). The algorithm analyzes each feature and output category as an independent variable and measures the similarity between two variables by using mutual information, as expressed by

$$\boldsymbol{MI(x,y) = \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy} \tag{1}$$

Where $p(x,y)$ represents the joint probabilistic density of x and y, and $p(x)$ and $p(y)$ represent the marginal probabilistic densities of x and y, respectively. Each time a feature is introduced to the mRMR process, the correlation between a feature set and a target must be determined. However, in feature selection, the combination of individual good features does not necessarily increase the performance of classifiers because the features may be highly correlated with each other and thus show redundancy. That is, the correlation between features and categorical variable are maximized, and the correlation between features are minimized. The formulas for maximizing correlation and minimizing redundancy are as follows:

$$\boldsymbol{\max D(S,c), D = \frac{1}{|S|} \sum\nolimits_{f_i \in S} MI(f_i, c)} \tag{2}$$

$$\boldsymbol{\min R(S), R = \frac{1}{|S|^2} \sum\nolimits_{f_i, f_j \in S} MI\left(f_i, f_j\right)} \tag{3}$$

Where S is the feature subset, $|S|$ is the number of features, f_i is the i-th feature, and c is the target category. Finally, the features are selected by maximizing the equation ϕ as follows:

$$\boldsymbol{max\, \phi(D,R), \phi = D - R} \tag{4}$$

However, it is not easy to obtain such feature subset as this problem is NP-hard. Accordingly, mRMR employs a heuristic way to complete this task. It repeatedly selects one feature with maximum relevance to target category and minimum redundancies to already-selected features. This procedure stops until all features have been selected. According to the selection order, features are sorted in a feature list. Evidently, features with high ranks are more important than those with low ranks.

We used the mRMR tool from http://home.penglab.com/proj/mRMR/ and used the default parameters for the analysis.

Incremental Feature Selection

Through mRMR method, we can obtain a feature list. However, it is still a problem which features should be selected. To determine the optimal features for one classification algorithm, the IFS method (Liu and Setiono, 1998) was employed.

IFS is a frequently used method for determining the ideal feature number for classification when combined with a classification algorithm (Liu and Setiono, 1998; Zhang et al., 2020; Zhang et al., 2021). Based on the feature list yielded by the mRMR method, it first builds a succession of feature subsets by one-step interval. The top feature in the list is included in the first feature subset, the top two features are included in the second feature subset, and so on. On each feature subset constructed, one classifier is generated based on the given classification algorithm and samples represented by the features in the subset. Such classifier is assessed through ten-fold cross-validation (Kohavi, 1995). The best classifier can be found, which was termed as the optimal classifier. The features used in such classifier were called optimal features and they comprised the optimal feature subset.

Synthetic Minority Oversampling Technique

As shown in **Table 1**, various cell types have different sample sizes. The sample size of hEnSC-derived hepatocytes was approximately 12 times that of EnSCs, and thus the sample size was highly unbalanced. This condition can lead to strong preferences in the training process, resulting in unreliable results. In the analysis of the effectiveness of each classifier, the synthetic minority oversampling technique (SMOTE) was used to lessen the impact of imbalance (Chawla et al., 2002; Ding et al., 2022; Pan et al., 2022; Zhou et al., 2022). The SMOTE implementation process consists of the following steps: *1*) randomly select one sample, say *x*, from a minority class; *2*) the *k* closest neighbors of *x* are obtained from all samples in the same minority class; *3*) sample $x_{i(nn)}$ is randomly selected from these *k* closest neighbors, and a random number ζ_1 between 0 and 1 is generated to synthesize a new sample x_{i1} with the following formula:

$$\boldsymbol{x_{i1}} = \boldsymbol{x_i} + \boldsymbol{\zeta}_1 \times \left(\boldsymbol{x_{i(nn)}} - \boldsymbol{x_i}\right) \tag{5}$$

This new sample is put into the minority class; *4*) above steps are repeated several times until the minority class has same number of samples in the majority class. In this project, the “SMOTE” tool from Weka was used. The new samples yielded by SMOTE were only used in the IFS method.

Classification Algorithm

Three efficient classification algorithms were used as candidates for the IFS method, which have been applied to tackle various biological and medical problems (Chen W. et al., 2021; Carlos et al., 2021; Liu et al., 2021; Li et al., 2022; Wu and Chen, 2022; Yang and Chen, 2022). They were briefly described as follows.

Random Forest

RF is an emerging and highly flexible machine learning algorithm that is widely used in biological data mining (Breiman, 2001). It is a typical type of ensemble classifier. The idea of an ensemble is to solve shortcomings inherent in a single model or a model with a certain set of parameters, and thus more models can be integrated, and limitations can be avoided. RFs are the products of the idea of ensemble, where many decision trees (DTs) are integrated into a forest for the prediction of a final outcome. Here, we called RF model from python’s scikit-learn package for classification. For convenience, we used default parameters to execute RF package. The number of DTs was 100.

k-Nearest Neighbor

KNN is the earliest collaborative filtering algorithm (Cover and Hart, 2003). The basic idea is to classify sample points that are close to one another into the same class. The KNN first determines a k-value which is used in selecting k-nearest samples in a specific point. Then, a selected distance is used in calculating the distance of the k-nearest samples to a specific point. Finally, a voting-based classification rule is used to determine the class to which the new sample belongs. We adopted the KNN model in scikit-learn for subsequent analysis. Default parameters were used, where the distance was defined as Minkowsk distance, *K* was set to one.

Decision Tree

DTs are machine learning algorithms with good interpretation, high training efficiency, and simple comprehension and frequently used in classification and feature selection (Safavian and Landgrebe, 1991). A DT splits in a recursive manner, resulting in a tree structure with nodes and directed edges. The classification of an instance is determined by sorting along the tree until it reaches a leaf node. In this study, we adopted DT implemented by the Scikit-learn package. It uses CART method with Gini index to expand the tree.

Performance Evaluation

The Matthews correlation coefficient (MCC) is a well-balanced indicator that may be used when the sample size is imbalanced (Matthews, 1975). It is used in measuring the binary classification problem and is more reliable than other measurements in biological data. Gorodkin proposed a widely used formulation of MCC in multi-class classification problems (Gorodkin, 2004). Such MCC can be determined using the formula below:

$$\begin{aligned}\boldsymbol{MCC} &= \frac{\boldsymbol{cov}(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{\boldsymbol{cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{cov}(\boldsymbol{Y},\boldsymbol{Y})}} \\ &= \frac{\frac{1}{K}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\boldsymbol{X_{nk}}-\bar{\boldsymbol{X}}_{\boldsymbol{k}}\right)\left(\boldsymbol{Y_{nk}}-\bar{\boldsymbol{Y}}_{\boldsymbol{k}}\right)}{\sqrt{\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\boldsymbol{X_{nk}}-\bar{\boldsymbol{X}}_{\boldsymbol{k}}\right)^2\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\boldsymbol{Y_{nk}}-\bar{\boldsymbol{Y}}_{\boldsymbol{k}}\right)^2}},\end{aligned} \tag{6}$$

Where X is the binary matrix into which one-hot encoding converts the predicted class of each sample, Y is another binary matrix into which one-hot encoding converts the real class of each sample, and $cov(X,Y)$ is the covariance of two matrices. The average of the kth column of matrices X and Y are represented by $\bar{X}_k$ and $\bar{Y}_k$, respectively. The elements in the n-th row and k-th column of the matrices X and Y are referred to as X_{nk} and Y_{nk}, respectively. The MCC range is [−1, 1], and 1 indicates that the forecasts are identical to actual outcomes, 0 indicates that the predictions are no difference from random, and

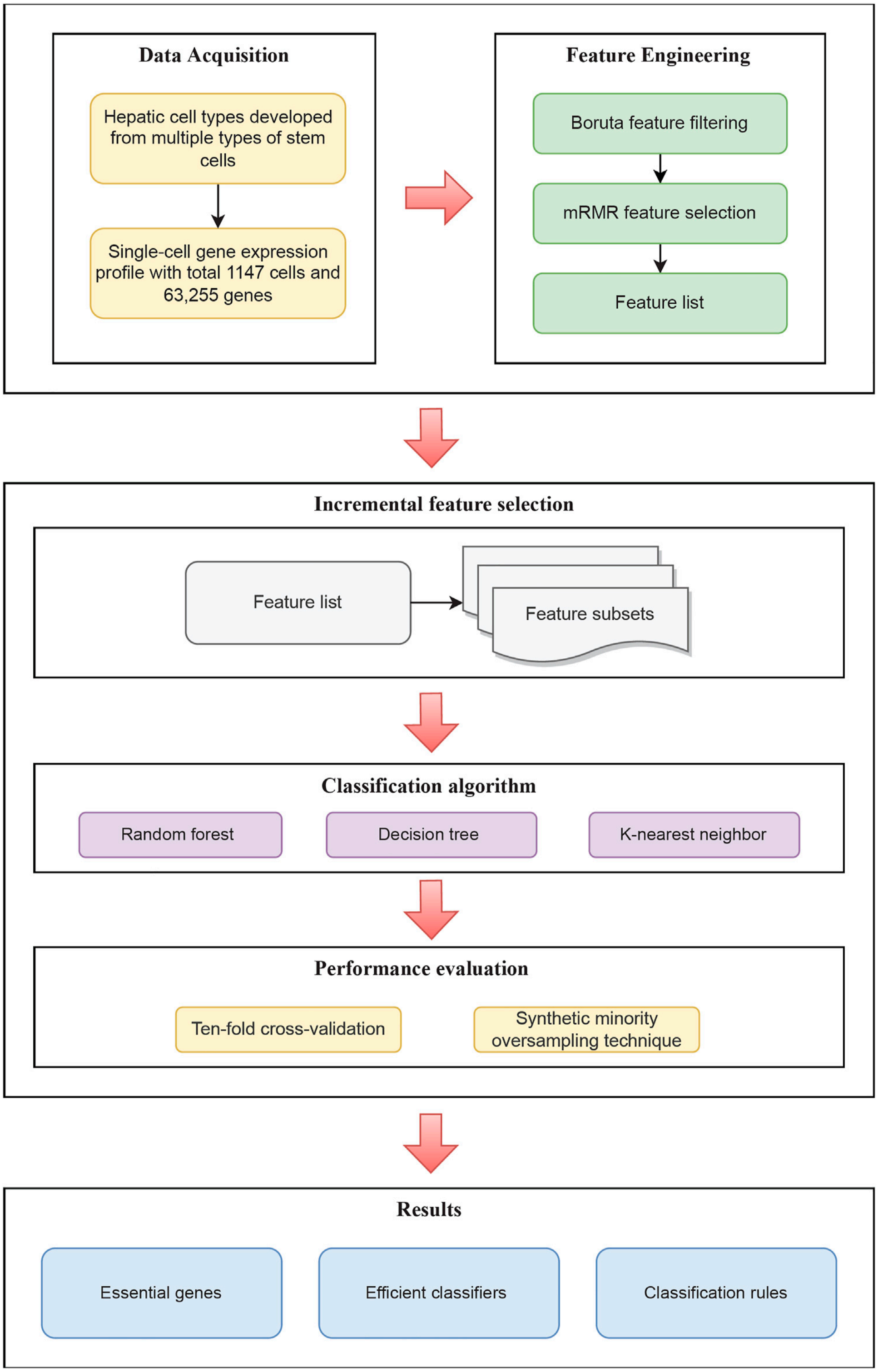


FIGURE 1 | Flow chart of the entire analysis process of this study. Single-cell RNA sequencing data acquired through the GEO database includes cells from 16 different hepatic cell types cultured *in vitro*. Following that, using feature selection methods, a sorted feature list is constructed. To recover efficient genes, develop effective classifiers, and construct classification rules, this list is partitioned into feature subsets and put into the three classification algorithms.

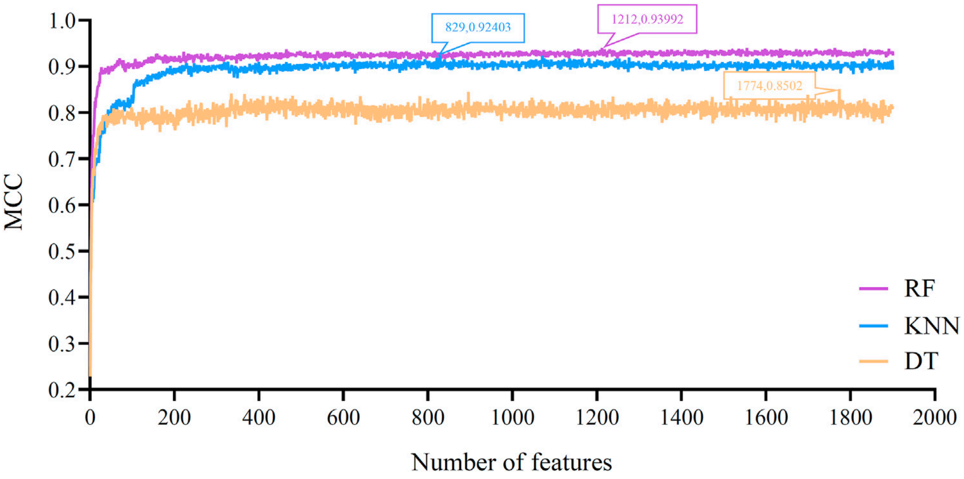


FIGURE 2 | IFS curves for evaluating the performance of the three classification algorithms under different feature subsets according to MCC. RF/KNN/DT reaches the maximum MCC value of 0.940/0.924/0.850 at the feature number of 1212/829/1774.

−1 indicates that the predictions are the polar opposites of the actual results.

In addition, some other widely used measurements for multi-class classification problems were also adopted in this study. They were overall accuracy (ACC) and individual accuracy on each class (cell type in this study). For the *i*-th class, its individual accuracy is defined as

$$ACC_i = \frac{n_i}{N_i}, \tag{7}$$

Where N_i stands for the number of samples in the *i*-th class and n_i is the number of correctly predicted samples in this class. As for ACC, it can be computed by

$$ACC = \frac{\sum_{i=1}^{16} n_i}{\sum_{i=1}^{16} N_i}, \tag{8}$$

Above measurements were provided as reference.

Functional Enrichment Analysis

We can get the optimal features for one classification algorithm using the IFS method. Functional enrichment analysis is critical for uncovering key pathways involved with the *in vitro* culture process and for unraveling the molecular processes of biomedicine. Thus, Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment studies were performed using the R package ClusterProfiler (Wu et al., 2021).

RESULTS

In the current research, we explored genes that characterize the process of hepatocyte culture and differentiation *in vitro* and created a series of rules for differentiating various hepatic cell types. The entire calculation process is shown in **Figure 1**. The outcomes of each step were discussed in full below.

Results of Boruta and mRMR Methods

We processed the original 63,255 features with the Boruta feature filtering approach. 1901 features were selected, which are listed in **Supplementary Table S1**. Subsequently, these features were analyzed by mRMR method, to obtain a list of features ranked by importance, which are also shown in **Supplementary Table S1**.

Results of the IFS Method

Based on the feature list obtained in **Results of Boruta and mRMR Methods** section, the IFS method was performed. It constructed 1,901 feature subsets with one step interval. On each feature subset, a classifier was built by applying one classification algorithm (RF, KNN or DT) to samples represented by features in this subset. Each classifier was evaluated by 10-fold cross-validation. The evaluation results, including measurements listed in **Performance Evaluation** section, are provided in **Supplementary Table S2**. To clear display the performance of one classification algorithm under different feature subsets, an IFS curve was plotted, as shown in **Figure 2**, which set MCC as *Y*-axis and number of features as *X*-axis. For RF, the highest MCC was 0.940, which was obtained by using top 1212 features in the list. Accordingly, the optimal RF classifier can be built with these features. The ACC of this classifier was 0.945, as listed in **Table 2**. Its detailed performance on 16 cell types (i.e., individual accuracies) is shown in **Figure 3**. It can be observed that several cell types were perfectly predicted. All these suggested the excellent high performance of the optimal RF classifier. As for another classification algorithm, KNN, its highest MCC was 0.924, which was produced by using top 829 features. With these features, the optimal KNN classifier was set up. Such classifier yielded the ACC of 0.930 (**Table 2**). The MCC and ACC were all lower than those of the optimal RF classifier. Its individual accuracies on 16 cell types were also generally lower than

TABLE 2 | 10-fold cross-validation performance of some key classifiers based on different classification algorithms.

Classification algorithm	Number of features	Overall accuracy (ACC)	Matthews correlation coefficient (MCC)
Random Forest	1212	0.945	0.940
Random Forest	222	0.937	0.931
k-Nearest Neighbor	829	0.930	0.924
Decision Tree	1774	0.863	0.850

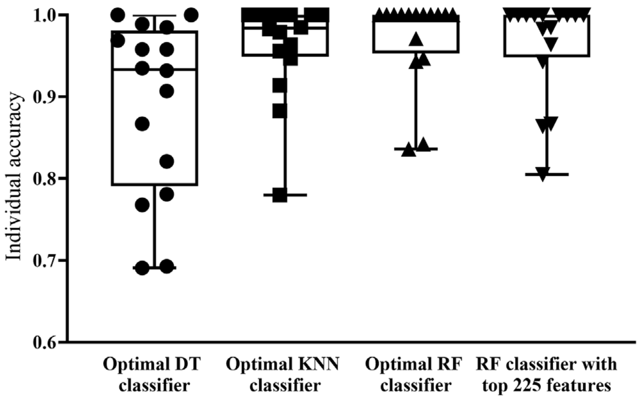


FIGURE 3 | Box plot to show performance of key classifiers on 16 cell types. RF and KNN classifiers have superior classification performance, with ACC reaching above 0.950 in most cell types. DT classifier has a weaker classification performance compared to the other three classifiers.

those of the optimal RF classifier, which can be observed from **Figure 3**.

With RF and KNN, the efficient classifiers can be built. However, they cannot provide useful clues to uncover the heterogeneity of the hepatocyte differentiation process *in vitro*. In view of this, this study further employed DT in the IFS method. The IFS curve of DT is also shown in **Figure 2**. When top 1,774 features were used, DT provided the highest MCC of 0.850. Likewise, the optimal DT classifier was constructed using these features. Its ACC was 0.863, as listed in **Table 2**. Evidently, such performance was much lower than that of the optimal RF/KNN classifier. Its performance on 16 cell types was also much lower than that of the other two optimal classifiers (**Figure 3**). Although the performance of the optimal DT classifier is much lower than the optimal KNN/RF classifier, it has its own merits, which would be given in **Classification Rules** section.

With the above arguments, we can find that the optimal RF classifier was best. Such classifier can be a useful tool to differentiate hepatic cell types cultured *in vitro*. However, the efficiency of this classifier was a problem because lots of features were used in this classifier. In view of this, we carefully checked the IFS results of RF and found that when top 222 features were adopted, RF can generate the MCC of 0.931. In this case, the ACC was 0.937 (**Table 2**). They were slightly lower than those of the optimal RF classifier. As for its individual accuracies, they were also a little lower than those of the optimal RF classifier, as shown in **Figure 3**. Furthermore, this RF classifier was superior to the optimal KNN and DT classifiers. Thus, it was more proper than the optimal RF classifier to be a tool for differentiating hepatic cell types cultured *in vitro*.

Classification Rules

By applying IFS method with DT to the *in vitro* cultured human hepatocyte single-cell RNA sequencing expression profiles, the optimal DT classifier was built. It used the top 1,774 features in the list. Although its performance was not very high, it can provide novel clues to uncover the heterogeneity of the hepatocyte differentiation process *in vitro*. With top 1,774 features, we applied DT on all cells, obtaining a large tree, from which 118 rules for classifying hepatic cell types were obtained. These rules are available in **Supplementary Table S3**. Each rule established a limit on the quantity of gene expression, indicating the relevance of high or low gene expression in distinguishing *in vitro* cultured cell types. Each cell type received at least one rules. **Figure 4** shows the number of rules for each cell type. The cell type "EnSC-derived hepatocyte" got the most rules (18), where four cell types only got one rule. In **Quantitative Rules for Stages of Liver Cells Differentiation and Specific Function Classification** section, a detailed analysis of these rules would be given.

Functional Enrichment Analyses

The IFS results showed that the optimal RF classifier provided the best classification performance. Such classifier used the top 1,212 features in the list, suggesting that these features greatly contributed to the model construction process for distinguishing the samples of different cell types and were directly or indirectly involved in the biological processes that distinguished these cells. To support this result, GO and KEGG pathway enrichment analysis was performed on the corresponding genes of these features by using ClusterProfiler (Wu et al., 2021) package in R. The FDR <0.05 criterion was used in filtering GO terms and KEGG pathways. **Supplementary Table S4** shows the results of GO and KEGG pathway enrichment analysis results. Then, we selected the top five GO terms in each GO group and KEGG pathways for visualization, as shown in **Figure 5**. Some terms, such as cell–substrate junction and cadherin binding, were linked to hepatocyte differentiation *in vitro* in these enrichment results. **Functional Enrichment Analysis of Optimum Genes** section presented a full analysis of the enrichment results.

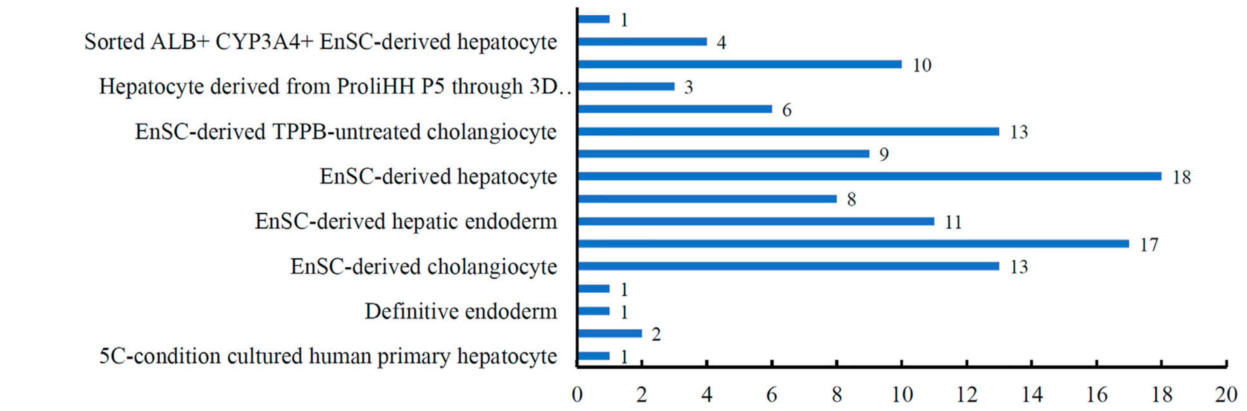


FIGURE 4 | Bar chart to show the number of rules for each cell type.

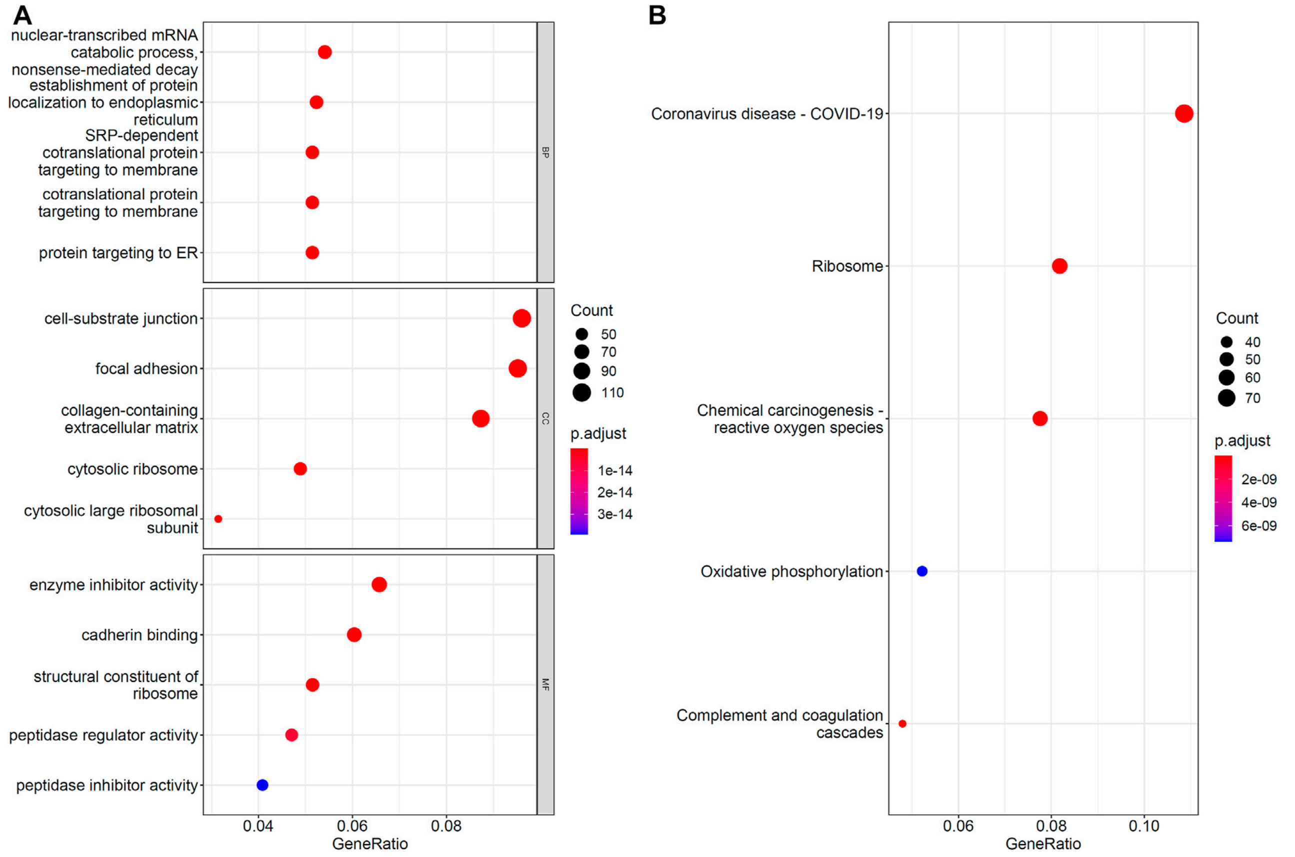


FIGURE 5 | Gene ontology and KEGG pathway enrichment analysis on optimal genes for RF. The FDR<0.05 criterion was used to filter GO terms and KEGG pathways. **(A)** The top five GO enrichment results for each GO group. **(B)** The top five KEGG pathway enrichment results.

DISCUSSION

We used advanced computational methods to identify qualitative features and quantitative rules for different stages of differentiation and specific functional populations of liver cells, which were cultured *in vitro*, at the single-cell level. The violin plot and heatmap were drawn using highly ranked genes to show expression patterns between different classes, which can be seen in **Figure 6**. These features play important roles in hepatocyte development, which also shows the accuracy of our analysis

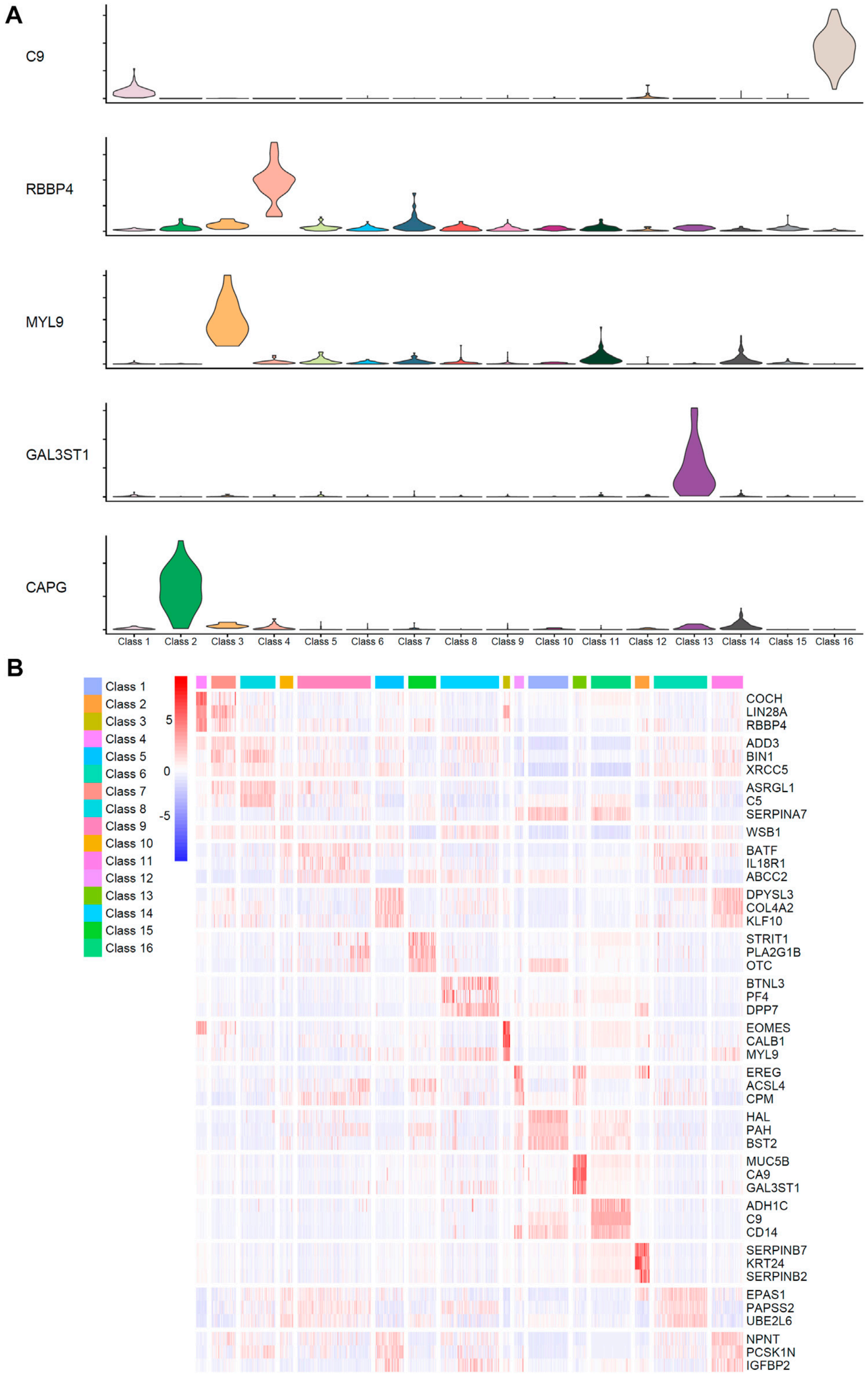


FIGURE 6 | Identified expression patterns of highly ranked genes among different classes. **(A)** The violin plot of five identified genes, *C9*, *RBBP4*, *MYL9*, *GAL3ST1*, and *CAPG*, which have significant high expression level in specific classes. **(B)** The heatmap of genes ranked high in the feature list. The corresponding cell types of Class 1–16 can be found in **Table 1**.

TABLE 3 | Important genes yielded by Boruta and mRMR methods.

Ensembl ID	Gene symbol	Description
ENSG00000034510	TMSB10	Thymosin Beta 10
ENSG00000172270	CD147/BSG	Basigin (Ok Blood Group)
ENSG00000106565	TMEM176B	Transmembrane Protein 176B
ENSG00000196262	PPIA	Peptidylprolyl Isomerase A
ENSG00000135404	CD63	CD63 Molecule

results. A detailed description of these features and rules can be seen below.

Optimal Features for Distinguishing Different Transplantable Liver Cells *In Vitro*

By the Boruta and mRMR methods, a feature list, indicating the importance of genes, were obtained. Here, we selected five genes with high ranks in the list for detailed analysis, which are listed in **Table 3**.

The first identified gene in the list was *TMSB10* (ENSG00000034510). *TMSB10* encodes the conserved small acid protein belonging to the beta-thymosin family, which functions in actin function during cell motility. TMSB10 expression is related to the development of several tissues (Bani-Yaghoub et al., 2001). Back in 1990, TMSB10 was found to be highly expressed during the human fetal brain period (Hall et al., 1990). In 2011, Fanni et al. found significant differences in the expression of TSM10 among the different stages of salivary gland organogenesis (Fanni et al., 2011). TSM10 is strongly expressed in the early stages of physiological development of human salivary glands (Nemolato et al., 2009; Fanni et al., 2011). Although no studies have directly shown that TSM10 plays an important role in liver formation and development, some studies implied the important role of TSM10 in embryonic development, revealing that TSM10 may be an important regulator in the differentiation of embryonic cells into hepatocytes.

CD147 (ENSG00000172270), also known as *basigin* (BSG), encodes a plasma membrane protein that plays important roles in life processes, such as embryo implantation and tumor progression. CD147 is one of the positive markers of a type of mesenchymal stem cells that are isolated from fetal liver (Zhao et al., 2004). This finding demonstrates the role of CD147 as a marker for identifying stem cells with high differentiation potential. It helped us select good starting cells during the *in vitro* culture of hepatocytes. CD147 regulates the production of MMP in hepatocytes and bile duct cells and reduces the degree of liver fibrosis (Calabro et al., 2014). CD147 expression affects carcinogenesis development by modulating the degree of cell differentiation in hepatocellular carcinoma (Wu et al., 2016). Through previous studies, we found that CD147 not only regulates terminally differentiated cells in the liver but also affects the differentiation process of the cells. Our method ranked it high in the list, indicating its importance in the differentiation and maturation of hepatocytes *in vitro*.

The next identified gene was *TMEM176B* (ENSG00000106565), which was first found in human lung fibroblasts (Lurton et al., 1999). TMEM176B was highly expressed in transplanted livers with recurrent hepatitis C virus, revealing its potential as a marker to distinguish abnormal reactions occurring after liver transplantation (Gehrau et al., 2011). Our study showed that TMEM176B was one of the efficient classification features, implying a specific pattern in TMEM176B expression among cell populations and further suggesting that diverse *in vitro* cultured cell populations have different adaptations for liver transplantation. In addition, TMEM176B regulates the maturation process of monocytes and dendritic cells in mice and humans (Condamine et al., 2010; Picotto et al., 2020). No direct evidence of the role of TMEM176A in hepatocyte differentiation was found, but the combination of previous and our studies revealed that TMEM176A potentially acts as a potential target for regulating hepatocyte maturation.

PPIA (ENSG00000196262), also known as *CYPA*, encodes a peptidyl-prolyl cis-trans isomerase that plays an important role in protein folding. It can act as a ligand to bind to CD147, thereby affecting intracellular physiological activities (Yurchenko et al., 2002). CD147, as described above, can affect the differentiation of cells within the liver. Therefore, PPIA is a potential target that influences hepatocyte differentiation. In addition, the inhibition of PPIA activity leads to the blocked polymerization of hensin in the extracellular matrix, thus preventing the full differentiation of epithelial cells (Peng et al., 2009). In 2005, CYPA was demonstrated to be involved in the early stages of neural differentiation (Urano et al., 2006). PPIA mediates many biological pathways, such as inflammation and apoptosis, but its function in the differentiation of embryonic hepatocytes *in vitro* has not been investigated. Previous studies and our studies showed its potential influence on functional cell differentiation.

The next identified gene was *CD63* (ENSG00000135404), which encodes a quadruple transmembrane protein localized on the surface of the cell membrane. This protein-mediated signal transduction event plays a role in the regulation of cell development, activation, growth, and motility (Pols and Klumperman, 2009). Exogenous TIMP-1 binds to CD63 and activates a series of pathways that ultimately mediate human hematopoietic stem or progenitor cells proliferation (Rossi et al., 2015). Thus, CD63 may act as a signaling initiator molecule that facilitates the proliferation and differentiation of stem cells *in vitro*, leading to the formation of cells with specific functions. In addition, CD63 interacts with ameloblastin in osteoblasts and promotes the interaction between CD63 and integrin β1, which ultimately promote osteogenic differentiation (Iizuka et al., 2011). CD63 is associated with cell differentiation in a variety of tissues and a potential target that influences the *in vitro* culture and differentiation of hepatocytes. Meanwhile, CD63 is one of the indicators for assessing liver regeneration and prognosis in patients with acute-on-chronic liver failure (Jiao et al., 2021). This result suggested that CD63 is critical to hepatocytes cultured *in vitro* and it may be directly related to the success of the subsequent transplantation of these cells into damaged livers.

Functional Enrichment Analysis of Optimum Genes

The IFS curve showed that the RF reached optimal performance in 1,212 features. We performed enrichment analysis on these 1,212 feature genes and filtered. The FDR was <0.05. The GO terms and KEGG pathways were directly or briefly involved in hepatocyte differentiation and functional formation, confirming the reliability of our selection method for the classification of hepatocytes at different stages of differentiation and cells with different functions. This result confirmed the validity of our selection method for the classification of hepatocytes at different stages of differentiation and different functions. We selected some of the top GO and KEGG enrichment results for detailed analysis.

In the biological process of GO enrichment results, GO: 0072599, which refers to the establishment of protein localization to the endoplasmic reticulum, displayed significant enrichment. Similar results were found for GO: 0070972, which refers to protein localization to the endoplasmic reticulum. During hepatocyte differentiation, changes in endoplasmic reticulum morphology and protein content in the microsomes on the endoplasmic reticulum were observed (Dallner et al., 1966; Kanamura et al., 1990). In addition, during liver development, endoplasmic reticulum processed large amounts of proteins and lipids to temporarily direct and perform proper functions (Hetz, 2012). In the cellular component of GO enrichment results, GO:0030055, which refers to the cell–substrate junction, showed high enrichment. Hepatocytes must interact with other cells and with a chemically complex substrates to maintain activity and function (Parsons-Wingerter and Saltzman, 1993). The biomechanical effects of cell–substrate interactions affect the differentiation of embryonic liver progenitor cells (Kourouklis et al., 2016). In the molecular function of GO enrichment results, GO:0045296, which refers to cadherin binding, was found to be significantly enriched. Calnexin-mediated intercellular contacts are essential to the *in vitro* maintenance of functioning hepatocytes (Semler et al., 2005). Moreover, the incorporation of E-calcineurin in cells containing appropriate substrates can maintain cell-specific functions in the liver and induce hepatocyte differentiation processes *in vitro* (Semler et al., 2005; Haque et al., 2011). Interestingly, in the KEGG enrichment analysis, we found hsa05171, which refers to the coronavirus disease (COVID-19), to be significantly enriched. Hepatocytes and cholangiocytes cultured *in vitro* are extremely permissive to SARS-CoV-2 infection (Yang et al., 2020). Hence, COVID19-related genes may be involved in the functional formation of hepatocytes and cholangiocytes *in vitro*.

Quantitative Rules for Stages of Liver Cells Differentiation and Specific Function Classification

In addition to qualitative features, we established a series of quantitative rules for distinguishing *in vitro* cultured liver cells. We classified these rules and cell clusters into two main categories. The first category included rules that distinguish specific cell clusters at different stages of hepatocyte differentiation *in vitro*. The second category included rules used in distinguishing specific hepatocyte clusters formed by the differentiation of different original cells *in vitro*. A detailed description of the rules can be found below.

First, the classification rules of six cell groups derived from the development of endodermal stem cells into hepatocytes and cholangiocytes were resolved. In developmental stages originating from endodermal stem cells, all the six cell types exhibited restricted SAA1, TMEM123, and CD36 expression. During the differentiation of stem cells into hepatocytes, SAA1 expression is upregulated in favor of liver metabolism, but the overexpression of SAA1 determines the development of inflammation (Shi et al., 2020; Choi et al., 2021). This finding was consistent with our results and showed the accuracy of our method. CD36 is involved in the metabolism of fat in hepatocytes, and high CD36 expression leads to fat accumulation and affects the normal functions of hepatocytes (Wilson et al., 2016; Li et al., 2019). PABPAC1 had high expression levels in Class 9 (hepatocyte) and Class 10 (immature hepatocyte) and low expression in other cells. The upregulated expression of PABPAC1 is associated with hepatocyte proliferation and growth (Hsieh et al., 2009). The classification rules for Class 4 (endoderm stem cell) and Class 7 (hepatic endoderm) showed a high degree of similarity, exhibiting the low expression of HAMP and SPTBN1 and high expression of APOE. HAMP, a protein specifically expressed in the liver, constitutes a major circulating regulator of iron uptake and distribution across tissues (Fang et al., 2020). Class 4 and Class 7 hepatocytes are cell populations in the early stages of differentiation and therefore have lower expression levels on hepatocyte-specific expressed genes. The inhibition of SPTBN1 in hepatocellular carcinoma cells increases the expression of stem cell markers, and this process is consistent with the less differentiated nature of these two types of cells (Zhi et al., 2015; Hu and Wu, 2021). APOE deficiency leads to liver senescence and is detrimental to hepatocyte differentiation (Bonomini et al., 2013). Thus, the high expression of APOE retains the strong differentiation abilities of Class 4 and Class 7 cells. In rule11, which was used in distinguishing Class 4 (endodermal stem cells), FOXH1 showed high expression levels. FOXH1 acts as a transcriptional co-activator and promotes the expression of MixL1, which plays an important role in the morphogenesis and endodermal differentiation of mouse embryos. In rule 7, which was used in distinguishing Class 5 (cholangiocyte), S100A6, GSTA1, and NOCA7 showed low expression levels, whereas QSOX1, BTG1 showed high expression levels. S100A6 plays a regulatory role in a variety of cell differentiation processes and has a low expression level in terminally differentiated cholangiocytes (Grahn et al., 2020). Given that high BTG1 expression inhibits cell proliferation and differentiation, cholangiocytes were presumed to have reached a stable state. Class 9 (hepatocyte) and Class 10 (immature hepatocyte) contained RPS27 in their classification rules, which had low expression in Class 9 and high expression in Class 10. High RPS27 expression has been reported in regenerating hepatocytes (Ganger et al., 2001). We hypothesized that RPS27 is a potential target for the transformation of immature hepatocytes into active mature hepatocytes.

The classification rules for the six classes of cell subtypes were resolved. These classes were hepatocytes obtained from the differentiation and development of three distinct original cells under different conditions. Class 1 included the primary hepatocytes maintained *in vitro* under 5C conditions, which brings the primary hepatocytes to a steady state by inhibiting a series of signaling pathways (Xiang et al., 2019). In rule 2, which was used for distinguishing Class 1, RAB5IF and CRIM1 showed low expression levels, whereas EMC7 showed high expression levels. In hepatocellular carcinoma, the RAB5I with low expression level binds to FLGR5, thereby inhibiting the proliferation of hepatocellular carcinoma cells (Koo et al., 2019). Inhibitory effect of RAB5I is similar to the inhibition of proliferation of primary hepatocytes under 5C conditions, indicating the accuracy of our method. CRIM1 is an important regulator of organ development and is highly expressed during differentiation (Iyer et al., 2016). primary cells maintained under 5C conditions are more stable and have lower differentiation indexes that those that are not, and CRIM1 has low expression level (Xiang et al., 2019). The function of EMC7 is currently undefined, but it is a potential target for maintaining the stability of primary hepatocytes *in vitro*. As for Class 16 (rule 3, uncultured adult human primary hepatocyte), SAA1 showed a high expression level in the classification rule. SAA1 encodes an acute phase protein that is highly expressed during tissue injury, inflammation, or infection (Li and Liao, 1999). Uncultured primary hepatocytes cannot maintain function *in vitro* for long periods of time. The cells may internally generate responses related to SAA1 function. In rule 16, which was used for distinguishing Class 13, XIST and CAT showed high expression levels. Highly expressed XIST binds miRNAs that inhibit cell differentiation, thereby promoting cell differentiation (Feng Y. et al., 2020). CAT is more highly expressed in immature cells than in mature cells, indicating that it is a maturation-associated gene (Tomisato et al., 2002). This finding is consistent with the characteristics of ProliHHs, which exhibits progenitor cell properties after multiple generations of culture (Zhang et al., 2018). As for Class 14 (rule 17, Human embryonic stem cell-derived hepatocyte-like cell), NRAGE and SPTBN1 showed high expression levels in the classification rule. The high expression of NRAGE facilitates the repair of homologous recombination and can make cells radioresistant by altering subcellular localization (Xue et al., 2010; Chang et al., 2018; Liu et al., 2020). The high expression of SPTBN1 can suppress inflammation in the liver (Lin et al., 2021). Our rule demonstrated the specificity of the function of hepatocytes differentiated from different original cells, proving the superiority of our method.

CONCLUSION

We used innovative and widely used computational approaches on single-cell RNA sequencing data to reveal the markers of various hepatic cell types. The results suggested the functional characteristics of each population of cells. The following three major aspects of our work are the end results of our efforts. The first is a list of genes that are potential targets for hepatocyte populations cultivated *in vitro* and related to specific markers. Some markers such as *CD147*, *PPIA*, *TMSB10*, *TMEM176B*, and *CD63* were identified, and these markers have been proven to be associated with hepatocyte differentiation and maturation *in vitro*. This aspect provides a theoretical foundation for understanding hepatocyte developmental processes and mechanisms and possible targets for clinical liver disease treatment. The second is the efficient classifier for determining the types of cells in the liver. The best random forest classifier with a 0.940 Matthews correlation coefficient had been constructed to distinguish different hepatic cell types. This classifier was trained on a vast amount of single-cell data and achieved outstanding classification results. The third aspect encompassed a set of classification rules as direct indicators of distinct cell types. The classification rules reveal the features of hepatic cell types at the level of quantitative gene expression, providing a theoretical foundation for the modification of hepatocytes to better function *in vivo*.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128060.

AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. LC performed the experiments. ZL and FH analyzed the results. ZL, FH, and LC wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200, XDA26040304), National Key R&D Program of China (2018YFC0910403), the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.916309/full#supplementary-material

REFERENCES

Aizarani, N., Saviano, A., SagarMailly, L., Mailly, L., Durand, S., Herman, J. S., et al. (2019). A Human Liver Cell Atlas Reveals Heterogeneity and Epithelial Progenitors. *Nature* 572, 199–204. doi:10.1038/s41586-019-1373-2

Bani-Yaghoub, M., Felker, J. M., Ozog, M. A., Bechberger, J. F., and Naus, C. C. G. (2001). Array Analysis of the Genes Regulated during Neuronal Differentiation of Human Embryonal Cells. *Biochem. Cell Biol.* 79, 387–398. doi:10.1139/o01-024

Bonomini, F., Rodella, L. F., Moghadasian, M., Lonati, C., and Rezzani, R. (2013). Apolipoprotein E Deficiency and a Mouse Model of Accelerated Liver Aging. *Biogerontology* 14, 209–220. doi:10.1007/s10522-013-9424-9

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Calabro, S. R., Maczurek, A. E., Morgan, A. J., Tu, T., Wen, V. W., Yee, C., et al. (2014). Hepatocyte Produced Matrix Metalloproteinases Are Regulated by CD147 in Liver Fibrogenesis. *PLoS One* 9, e90571. doi:10.1371/journal.pone.0090571

Carlos, M., Zoran, K., and Juan, S. (2021). Predicting Non-deposition Sediment Transport in Sewer Pipes Using Random Forest. *Water Res.* 189, 116639. doi:10.1016/j.watres.2020.116639

Chang, X., Xue, X., Zhang, Y., Zhang, G., Zhou, H., Yang, Y., et al. (2018). The Role of NRAGE Subcellular Location and Epithelial-Mesenchymal Transition on Radiation Resistance of Esophageal Carcinoma Cell. *J. Cancer Res. Ther.* 14, 46–51. doi:10.4103/jcrt.JCRT_687_17

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *jair* 16, 321–357. doi:10.1613/jair.953

Chen, L., Li, Z., Zeng, T., Zhang, Y. H., Feng, K., Huang, T., et al. (2021a). Identifying COVID-19-specific Transcriptomic Biomarkers with Machine Learning Methods. *Biomed. Res. Int.* 2021, 9939134. doi:10.1155/2021/9939134

Chen, L., Li, Z., Zhang, S., Zhang, Y. H., Huang, T., and Cai, Y. D. (2022). Predicting RNA 5-methylcytosine Sites by Using Essential Sequence Features and Distributions. *Biomed. Res. Int.* 2022, 4035462. doi:10.1155/2022/4035462

Chen, W., Chen, L., and Dai, Q. (2021b). iMPT-FDNPL: Identification of Membrane Protein Types with Functional Domains and a Natural Language Processing Approach. *Comput. Math. Methods Med.* 2021, 7681497. doi:10.1155/2021/7681497

Choi, M., Park, S., Yi, J. K., Kwon, W., Jang, S., Kim, S.-Y., et al. (2021). Overexpression of Hepatic Serum Amyloid A1 in Mice Increases IL-17-producing Innate Immune Cells and Decreases Bone Density. *J. Biol. Chem.* 296, 100595. doi:10.1016/j.jbc.2021.100595

Condamine, T., Le Texier, L., Howie, D., Lavault, A., Hill, M., Halary, F., et al. (2010). Tmem176B and Tmem176A Are Associated with the Immature State of Dendritic Cells. *J. Leukoc. Biol.* 88, 507–515. doi:10.1189/jlb.1109738

Cover, T., and Hart, P. (2003). Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* 13, 21–27.

Dallner, G., Siekevitz, P., and Palade, G. E. (1966). Biogenesis of Endoplasmic Reticulum Membranes. *J. Cell Biol.* 30, 97–117. doi:10.1083/jcb.30.1.97

Ding, S., Li, H., Zhang, Y.-H., Zhou, X., Feng, K., Li, Z., et al. (2021). Identification of Pan-Cancer Biomarkers Based on the Gene Expression Profiles of Cancer Cell Lines. *Front. Cell Dev. Biol.* 9, 781285. doi:10.3389/fcell.2021.781285

Ding, S., Wang, D., Zhou, X., Chen, L., Feng, K., Xu, X., et al. (2022). Predicting Heart Cell Types by Using Transcriptome Profiles and a Machine Learning Method. *Life* 12, 228. doi:10.3390/life12020228

Du, X., Chang, S., Guo, W., Zhang, S., and Chen, Z. K. (2020). Progress in Liver Transplant Tolerance and Tolerance-Inducing Cellular Therapies. *Front. Immunol.* 11, 1326. doi:10.3389/fimmu.2020.01326

Fang, Z., Zhu, Z., Zhang, H., Peng, Y., Liu, J., Lu, H., et al. (2020). GDF11 Contributes to Hepatic Hepcidin (HAMP) Inhibition through SMURF1-mediated BMP-SMAD Signalling Suppression. *Br. J. Haematol.* 188, 321–331. doi:10.1111/bjh.16156

Fanni, D., Gerosa, C., Nemolato, S., Locci, A., Marinelli, V., Cabras, T., et al. (2011). Thymosin Beta 10 Expression in Developing Human Salivary Glands. *Early Hum. Dev.* 87, 779–783. doi:10.1016/j.earlhumdev.2011.06.001

Feng, S., Wu, J., Qiu, W.-L., Yang, L., Deng, X., Zhou, Y., et al. (2020a). Large-scale Generation of Functional and Transplantable Hepatocytes and Cholangiocytes from Human Endoderm Stem Cells. *Cell Rep.* 33, 108455. doi:10.1016/j.celrep.2020.108455

Feng, Y., Wan, P., and Yin, L. (2020b). Long Noncoding RNA X-Inactive Specific Transcript (XIST) Promotes Osteogenic Differentiation of Periodontal Ligament Stem Cells by Sponging MicroRNA-214-3p. *Med. Sci. Monit.* 26, e918932. doi:10.12659/MSM.918932

Ganger, D. R., Hamilton, P. D., Klos, D. J., Jakate, S., Mcchesney, L., and Fernandez-Pol, J. A. (2001). Differential Expression of metallopanstimulin/S27 Ribosomal Protein in Hepatic Regeneration and Neoplasia. *Cancer Detect Prev.* 25, 231–236.

Gehrau, R., Maluf, D., Archer, K., Stravitz, R., Suh, J., Le, N., et al. (2011). Molecular Pathways Differentiate Hepatitis C Virus (HCV) Recurrence from Acute Cellular Rejection in HCV Liver Recipients. *Mol. Med.* 17, 824–833. doi:10.2119/molmed.2011.00072

Gorodkin, J. (2004). Comparing Two K-Category Assignments by a K-Category Correlation Coefficient. *Comput. Biol. Chem.* 28, 367–374. doi:10.1016/j.compbiolchem.2004.09.006

Grahn, T. H. M., Niroula, A., Végvári, Á., Oburoglu, L., Pertesi, M., Warsi, S., et al. (2020). S100A6 Is a Critical Regulator of Hematopoietic Stem Cells. *Leukemia* 34, 3323–3337. doi:10.1038/s41375-020-0901-2

Guo, R., Xu, X., Lu, Y., and Xie, X. (2017). Physiological Oxygen Tension Reduces Hepatocyte Dedifferentiation in In Vitro Culture. *Sci. Rep.* 7, 5923. doi:10.1038/s41598-017-06433-3

Hall, A. K., Hempstead, J., and Morgan, J. I. (1990). Thymosin β10 Levels in Developing Human Brain and its Regulation by Retinoic Acid in the HTB-10 Neuroblastoma. *Mol. Brain Res.* 8, 129–135. doi:10.1016/0169-328x(90)90057-k

Hanchuan Peng, H. C., Fuhui Long, F. H., and Ding, C. (2005). Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi:10.1109/tpami.2005.159

Haque, A., Hexig, B., Meng, Q., Hossain, S., Nagaoka, M., and Akaike, T. (2011). The Effect of Recombinant E-Cadherin Substratum on the Differentiation of Endoderm-Derived Hepatocyte-like Cells from Embryonic Stem Cells. *Biomaterials* 32, 2032–2042. doi:10.1016/j.biomaterials.2010.11.045

Hetz, C. (2012). The Unfolded Protein Response: Controlling Cell Fate Decisions under ER Stress and beyond. *Nat. Rev. Mol. Cell Biol.* 13, 89–102. doi:10.1038/nrm3270

Hsieh, H.-C., Chen, Y.-T., Li, J.-M., Chou, T.-Y., Chang, M.-F., Huang, S.-C., et al. (2009). Protein Profilings in Mouse Liver Regeneration after Partial Hepatectomy Using iTRAQ Technology. *J. Proteome Res.* 8, 1004–1013. doi:10.1021/pr800696m

Hu, C., and Li, L. (2015). In Vitro culture of Isolated Primary Hepatocytes and Stem Cell-Derived Hepatocyte-like Cells for Liver Regeneration. *Protein Cell* 6, 562–574. doi:10.1007/s13238-015-0180-2

Hu, L., and Wu, C. (2021). In Silico analysis Suggests Disruption of Interactions between HAMP from Hepatocytes and SLC40A1 from Macrophages in Hepatocellular Carcinoma. *BMC Med. Genomics* 14, 128. doi:10.1186/s12920-021-00977-0

Iansante, V., Mitry, R. R., Filippi, C., Fitzpatrick, E., and Dhawan, A. (2018). Human Hepatocyte Transplantation for Liver Disease: Current Status and Future Perspectives. *Pediatr. Res.* 83, 232–240. doi:10.1038/pr.2017.284

Iizuka, S., Kudo, Y., Yoshida, M., Tsunematsu, T., Yoshiko, Y., Uchida, T., et al. (2011). Ameloblastin Regulates Osteogenic Differentiation by Inhibiting Src Kinase via Cross Talk between Integrin β1 and CD63. *Mol. Cell Biol.* 31, 783–792. doi:10.1128/mcb.00912-10

Iyer, S., Pennisi, D. J., and Piper, M. (2016). Crim1-, a Regulator of Developmental Organogenesis. *Histol. Histopathol.* 31, 1049–1057. doi:10.14670/HH-11-766

Jiao, Y., Lu, W., Xu, P., Shi, H., Chen, D., Chen, Y., et al. (2021). Hepatocyte-derived Exosome May Be as a Biomarker of Liver Regeneration and Prognostic Valuation in Patients with Acute-On-Chronic Liver Failure. *Hepatol. Int.* 15, 957–969. doi:10.1007/s12072-021-10217-3

Kanamura, S., Kanai, K., and Watanabe, J. (1990). Fine Structure and Function of Hepatocytes during Development. *J. Elec. Microsc. Tech.* 14, 92–105. doi:10.1002/jemt.1060140204

Kawahara, T., Yagita, H., Kasai, S., Sawa, M., Kato, K., Okumura, K., et al. (1998). Allogeneic Hepatocyte Transplantation: Contribution of Fas-Fas Ligand Interaction to Allogeneic Hepatocyte Rejection. *J. Gastroenterol. Hepatol.* 13, S119–s123. doi:10.1111/jgh.1998.13.s1.119

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2* (Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.).

Koo, J. I., Lee, H. J., Jung, J. H., Im, E., Kim, J. H., Shin, N., et al. (2019). The Pivotal Role of Long Noncoding RNA RAB5IF in the Proliferation of Hepatocellular Carcinoma via LGR5 Mediated β-Catenin and C-Myc Signaling. *Biomolecules* 9, 718. doi:10.3390/biom9110718

Kourouklis, A. P., Kaylan, K. B., and Underhill, G. H. (2016). Substrate Stiffness and Matrix Composition Coordinately Control the Differentiation of Liver Progenitor Cells. *Biomaterials* 99, 82–94. doi:10.1016/j.biomaterials.2016.05.016

Kursa, M. B., and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *J. Stat. Softw.* 36, 1–13. doi:10.18637/jss.v036.i11

Lauschke, V. M., Shafagh, R. Z., Hendriks, D. F. G., and Ingelman-Sundberg, M. (2019). 3D Primary Hepatocyte Culture Systems for Analyses of Liver Diseases, Drug Metabolism, and Toxicity: Emerging Culture Paradigms and Applications. *Biotechnol. J.* 14, e1800347. doi:10.1002/biot.201800347

Li, L., and Liao, W. S.-L. (1999). An Upstream Repressor Element that Contributes to Hepatocyte-specific Expression of the Rat Serum Amyloid A1 Gene. *Biochem. Biophysical Res. Commun.* 264, 395–403. doi:10.1006/bbrc.1999.1527

Li, X., Lu, L., Lu, L., and Chen, L. (2022). Identification of Protein Functions in Mouse with a Label Space Partition Method. *Mbe* 19, 3820–3842. doi:10.3934/mbe.2022176

Li, Y., Yang, P., Zhao, L., Chen, Y., Zhang, X., Zeng, S., et al. (2019). CD36 Plays a Negative Role in the Regulation of Lipophagy in Hepatocytes through an AMPK-dependent Pathway. *J. Lipid Res.* 60, 844–855. doi:10.1194/jlr.m090969

Lin, L., Chen, S., Wang, H., Gao, B., Kallakury, B., Bhuvaneshwar, K., et al. (2021). SPTBN1 Inhibits Inflammatory Responses and Hepatocarcinogenesis via the Stabilization of SOCS1 and Downregulation of P65 in Hepatocellular Carcinoma. *Theranostics* 11, 4232–4250. doi:10.7150/thno.49819

Liu, H., Hu, B., Chen, L., and Lu, L. (2021). Identifying Protein Subcellular Location with Embedding Features Learned from Networks. *Cp* 18, 646–660. doi:10.2174/1570164617999201124142950

Liu, H., and Setiono, R. (1998). Incremental Feature Selection. *Appl. Intell.* 9, 217–230. doi:10.1023/a:1008363719778

Liu, L., Cui, Z., Zhang, J., Wang, J., Gu, S., Ma, J., et al. (2020). Knockdown of NRAGE Impairs Homologous Recombination Repair and Sensitizes Hepatoblastoma Cells to Ionizing Radiation. *Cancer Biotherapy Radiopharm.* 35, 41–49. doi:10.1089/cbr.2019.2968

Logan, D. J., Shan, J., Bhatia, S. N., and Carpenter, A. E. (2016). Quantifying Co-cultured Cell Phenotypes in High-Throughput Using Pixel-Based Classification. *Methods* 96, 6–11. doi:10.1016/j.ymeth.2015.12.002

Lurton, J., Rose, T. M., Raghu, G., and Narayanan, A. S. (1999). Isolation of a Gene Product Expressed by a Subpopulation of Human Lung Fibroblasts by Differential Display. *Am. J. Respir. Cell Mol. Biol.* 20, 327–331. doi:10.1165/ajrcmb.20.2.3368

Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica Biophysica Acta (BBA) - Protein Struct.* 405, 442–451. doi:10.1016/0005-2795(75)90109-9

Nemolato, S., Messana, I., Cabras, T., Manconi, B., Inzitari, R., Fanali, C., et al. (2009). Thymosin β4 and β10 Levels in Pre-term Newborn Oral Cavity and Foetal Salivary Glands Evidence a Switch of Secretion during Foetal Development. *PLoS One* 4, e5109. doi:10.1371/journal.pone.0005109

Pan, X., Chen, L., Liu, I., Niu, Z., Huang, T., and Cai, Y. D. (2022). Identifying Protein Subcellular Locations with Embeddings-Based Node2loc. *IEEE/ACM Trans. Comput. Biol. Bioinform* 19, 666–675.

Parsons-Wingerter, P. A., and Saltzman, W. M. (1993). Growth versus Function in the Three-Dimensional Culture of Single and Aggregated Hepatocytes within Collagen Gels. *Biotechnol. Prog.* 9, 600–607. doi:10.1021/bp00024a006

Peng, H., Vijayakumar, S., Schiene-Fischer, C., Li, H., Purkerson, J. M., Malesevic, M., et al. (2009). Secreted Cyclophilin A, a Peptidylprolyl Cis-Trans Isomerase, Mediates Matrix Assembly of Hensin, a Protein Implicated in Epithelial Differentiation. *J. Biol. Chem.* 284, 6465–6475. doi:10.1074/jbc.m808964200

Picotto, G., Morse, L. R., Nguyen, N., Saltzman, J., and Battaglino, R. (2020). TMEM176A and TMEM176B Are Candidate Regulators of Inhibition of Dendritic Cell Maturation and Function after Chronic Spinal Cord Injury. *J. Neurotrauma* 37, 528–533. doi:10.1089/neu.2019.6498

Pols, M. S., and Klumperman, J. (2009). Trafficking and Function of the Tetraspanin CD63. *Exp. Cell Res.* 315, 1584–1592. doi:10.1016/j.yexcr.2008.09.020

Rossi, L., Forte, D., Migliardi, G., Salvestrini, V., Buzzi, M., Ricciardi, M. R., et al. (2015). The Tissue Inhibitor of Metalloproteinases 1 Increases the Clonogenic Efficiency of Human Hematopoietic Progenitor Cells through CD63/PI3K/Akt Signaling. *Exp. Hematol.* 43, 974–985. e971. doi:10.1016/j.exphem.2015.07.003

Safavian, S. R., and Landgrebe, D. (1991). A Survey of Decision Tree Classifier Methodology. *IEEE Trans. Syst. Man. Cybern.* 21, 660–674. doi:10.1109/21.97458

Semler, E. J., Dasgupta, A., and Moghe, P. V. (2005). Cytomimetic Engineering of Hepatocyte Morphogenesis and Function by Substrate-Based Presentation of Acellular E-Cadherin. *Tissue Eng.* 11, 734–750. doi:10.1089/ten.2005.11.734

Shi, D., Xin, J., Lu, Y., Ding, W., Jiang, J., Zhou, Q., et al. (2020). Transcriptome Profiling Reveals Distinct Phenotype of Human Bone Marrow Mesenchymal Stem Cell-Derived Hepatocyte-like Cells. *Int. J. Med. Sci.* 17, 263–273. doi:10.7150/ijms.36255

Thomson, A. W., Vionnet, J., and Sanchez-Fueyo, A. (2020). Understanding, Predicting and Achieving Liver Transplant Tolerance: from Bench to Bedside. *Nat. Rev. Gastroenterol. Hepatol.* 17, 719–739. doi:10.1038/s41575-020-0334-4

Tomisato, W., Hoshino, T., Tsutsumi, S., Tsuchiya, T., and Mizushima, T. (2002). Maturation-associated Increase in Sensitivity of Cultured guinea Pig Gastric Pit Cells to Hydrogen Peroxide. *Dig. Dis. Sci.* 47, 2125–2133. doi:10.1023/a:1019653719397

Urano, Y., Iiduka, M., Sugiyama, A., Akiyama, H., Uzawa, K., Matsumoto, G., et al. (2006). Involvement of the Mouse Prp19 Gene in Neuronal/astroglial Cell Fate Decisions. *J. Biol. Chem.* 281, 7498–7514. doi:10.1074/jbc.m510881200

Wilson, C. G., Tran, J. L., Erion, D. M., Vera, N. B., Febbraio, M., and Weiss, E. J. (2016). Hepatocyte-Specific Disruption of CD36 Attenuates Fatty Liver and Improves Insulin Sensitivity in HFD-Fed Mice. *Endocrinology* 157, 570–585. doi:10.1210/en.2015-1866

Wu, J., Lu, M., Li, Y., Shang, Y.-K., Wang, S.-J., Meng, Y., et al. (2016). Regulation of a TGF-B1-Cd147 Self-Sustaining Network in the Differentiation Plasticity of Hepatocellular Carcinoma Cells. *Oncogene* 35, 5468–5479. doi:10.1038/onc.2016.89

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation* 2, 100141. doi:10.1016/j.xinn.2021.100141

Wu, Z., and Chen, L. (2022). Similarity-based Method with Multiple-Feature Sampling for Predicting Drug Side Effects. *Comput. Math. Methods Med.* 2022, 9547317. doi:10.1155/2022/9547317

Xiang, C., Du, Y., Meng, G., Soon Yi, L., Sun, S., Song, N., et al. (2019). Long-term Functional Maintenance of Primary Human Hepatocytes In Vitro. *Science* 364, 399–402. doi:10.1126/science.aau7307

Xiao, J., Wang, F., Wong, N.-K., He, J., Zhang, R., Sun, R., et al. (2019). Global Liver Disease Burdens and Research Trends: Analysis from a Chinese Perspective. *J. Hepatology* 71, 212–221. doi:10.1016/j.jhep.2019.03.004

Xue, X.-Y., Liu, Z.-H., Jing, F.-M., Li, Y.-G., Liu, H.-Z., and Gao, X.-S. (2010). Relationship between NRAGE and the Radioresistance of Esophageal Carcinoma Cell Line TE13R120. *Chin. J. Cancer* 29, 900–906. doi:10.5732/cjc.010.10141

Yang, L., Han, Y., Nilsson-Payant, B. E., Gupta, V., Wang, P., Duan, X., et al. (2020). A Human Pluripotent Stem Cell-Based Platform to Study SARS-CoV-2 Tropism and Model Virus Infection in Human Cells and Organoids. *Cell Stem Cell* 27, 125–136. e127. doi:10.1016/j.stem.2020.06.015

Yang, Y., and Chen, L. (2022). Identification of Drug-Disease Associations by Using Multiple Drug and Disease Networks. *Cbio* 17, 48–59. doi:10.2174/1574893616666210825115406

Yurchenko, V., Zybarth, G., O'connor, M., Dai, W. W., Franchin, G., Hao, T., et al. (2002). Active Site Residues of Cyclophilin A Are Crucial for its Signaling Activity via CD147. *J. Biol. Chem.* 277, 22959–22965. doi:10.1074/jbc.m201593200

Zhang, K., Zhang, L., Liu, W., Ma, X., Cen, J., Sun, Z., et al. (2018). In Vitro Expansion of Primary Human Hepatocytes with Efficient Liver Repopulation Capacity. *Cell Stem Cell* 23, 806–819. e804. doi:10.1016/j.stem.2018.10.018

Zhang, Y.-H., Guo, W., Zeng, T., Zhang, S., Chen, L., Gamarra, M., et al. (2021). Identification of Microbiota Biomarkers with Orthologous Gene Annotation for Type 2 Diabetes. *Front. Microbiol.* 12, 711244. doi:10.3389/fmicb.2021.711244

Zhang, Y.-H., Li, Z., Zeng, T., Pan, X., Chen, L., Liu, D., et al. (2020). Distinguishing Glioblastoma Subtypes by Methylation Signatures. *Front. Genet.* 11, 604336. doi:10.3389/fgene.2020.604336

Zhao, X. Z., Wei, L., Han, M., and Li, L. S. (2004). Isolation, Culture and Multipotent Differentiation of Mesenchymal Stem Cells from Human Fetal Livers. *Zhonghua Gan Zang Bing Za Zhi* 12, 711–713.

Zhi, X., Lin, L., Yang, S., Bhuvaneshwar, K., Wang, H., Gusev, Y., et al. (2015). Bii-Spectrin (SPTBN1) Suppresses Progression of Hepatocellular Carcinoma and Wnt Signaling by Regulation of Wnt Inhibitor Kallistatin. *Hepatology* 61, 598–612. doi:10.1002/hep.27558

Zhou, X., Ding, S., Wang, D., Chen, L., Feng, K., Huang, T., et al. (2022). Identification of Cell Markers and Their Expression Patterns in Skin Based on Single-Cell RNA-Sequencing profiles. *Life*. 12. 550. doi:10.3390/life12040550

Zhu, L., Yang, X., Zhu, R., and Yu, L. (2020). Identifying Discriminative Biological Function Features and Rules for Cancer-Related Long Non-coding RNAs. *Front. Genet.* 11, 598773. doi:10.3389/fgene.2020.598773

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 | Frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 02 June 2022
doi: 10.3389/fbioe.2022.890901

Identification of Type 2 Diabetes Biomarkers From Mixed Single-Cell Sequencing Data With Feature Selection Methods

Zhandong Li[1†], *Xiaoyong Pan*[2†] *and Yu-Dong Cai*[3]*

[1]College of Biological and Food Engineering, Jilin Engineering Normal University, Changchun, China, [2]Key Laboratory of System Control and Information Processing, Institute of Image Processing and Pattern Recognition, Ministry of Education of China, Shanghai Jiao Tong University, Shanghai, China, [3]School of Life Sciences, Shanghai University, Shanghai, China

Diabetes is the most common disease and a major threat to human health. Type 2 diabetes (T2D) makes up about 90% of all cases. With the development of high-throughput sequencing technologies, more and more fundamental pathogenesis of T2D at genetic and transcriptomic levels has been revealed. The recent single-cell sequencing can further reveal the cellular heterogenicity of complex diseases in an unprecedented way. With the expectation on the molecular essence of T2D across multiple cell types, we investigated the expression profiling of more than 1,600 single cells (949 cells from T2D patients and 651 cells from normal controls) and identified the differential expression profiling and characteristics at the transcriptomics level that can distinguish such two groups of cells at the single-cell level. The expression profile was analyzed by several machine learning algorithms, including Monte Carlo feature selection, support vector machine, and repeated incremental pruning to produce error reduction (RIPPER). On one hand, some T2D-associated genes (MTND4P24, MTND2P28, and LOC100128906) were discovered. On the other hand, we revealed novel potential pathogenic mechanisms in a rule manner. They are induced by newly recognized genes and neglected by traditional bulk sequencing techniques. Particularly, the newly identified T2D genes were shown to follow specific quantitative rules with diabetes prediction potentials, and such rules further indicated several potential functional crosstalks involved in T2D.

Keywords: type 2 diabetes, single-cell sequencing, Monte Carlo feature selection, support vector machine, RIPPER

OPEN ACCESS

Edited by:
Jiaofang Shao,
Nanjing Medical University, China

Reviewed by:
Bo Zhou,
Shanghai University of Medicine and Health Sciences, China
Gele Aori,
University of Toyama, Japan

Correspondence:
Yu-Dong Cai
cai_yud@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *07 March 2022*
Accepted: *04 April 2022*
Published: *02 June 2022*

Citation:
Li Z, Pan X and Cai Y-D (2022) Identification of Type 2 Diabetes Biomarkers From Mixed Single-Cell Sequencing Data With Feature Selection Methods.
Front. Bioeng. Biotechnol. 10:890901.
doi: 10.3389/fbioe.2022.890901

1 INTRODUCTION

Diabetes mellitus (DM) turns out to be the general term describing metabolic disorders with high blood sugar levels as typical symptoms (Tseng et al., 2012; Tao et al., 2015). Due to either lack of insulin or pathogenic insulin reactive responses, diabetes can be divided into three groups: type 1 DM with low insulin production, type 2 DM with insulin resistance, and gestational diabetes with high blood sugar levels induced by diabetes recurrence during pregnancy (American Diabetes Association, 2014). According to the epidemiologic statistics data in 2015, more than four hundred million people suffered from diabetes, and about five million people died from such disease all over the world (Gao et al., 2016; Disease and Injury Incidence and Prevalence Collaborators, 2017; Global Burden of Disease Cancer Collaboration et al., 2017). Particularly, type 2 diabetes (T2D) makes up about 90% of all cases (392 million) and is the primary subtype of diabetes (Disease and Injury Incidence and Prevalence Collaborators, 2017; Global Burden of Disease Cancer Collaboration et al., 2017), indicating such kind of disease is one of the major threats to human health.

Different from type 1 DM and gestational diabetes, the major pathogenesis of type 2 DM is insulin resistance and beta-cell dysfunction accompanied with insufficient insulin secretion (Pandey et al., 2015), where insulin resistance is generally defined as dysfunctional insulin-mediated glucose clearance (Yabe et al., 2015). During the pathogenesis of type 2 DM, the typical insulin-associated biological processes and action cascade are usually disturbed by either intracellular signals or extra interferences, including serine phosphorylation of IRS-1, excess glucosamine, mitochondria defects, FA (fatty acid)-induced insulin dysfunction, and alternate fatty acid effects (Taylor, 2013; Eckardt et al., 2014; Pandey et al., 2015). Early in 1997, Boden (1997) has already demonstrated the significance of fatty acids in diabetes. Further similarly in the same year, functional signaling molecules IRS-1 and IRS-2 were confirmed by Zick (2001), revealing the initial biological foundations for diabetes. Apart from such complicated pathogenesis associated with insulin resistance, beta-cell dysfunction has also been widely identified in type 2 DM patients as the other etiological factor. Similar to insulin resistance, such pathogenesis also has various potential mechanisms, including glucose toxicity, beta-cell exhaustion, impaired proinsulin biosynthesis, and lipo-toxicity (Ferrannini, 2009). In 2003, Kahn (2003) demonstrated the specific contribution of both insulin resistance and beta-cell dysfunction to the pathogenesis of diabetes, laying a foundation for the basic pathological mechanisms of such disease. Different from the downstream mechanisms of two major pathogeneses, such pathogenic mechanisms can be both attributed to either genetic predisposition or environmental interferences (Andersen et al., 2016; Stancakova and Laakso, 2016). They would be involved in the progressive dysfunction of pancreatic islet alpha and beta cells, so that, the pancreatic islet cells actually have specific roles in the initiation and progression of type 2 DM.

Traditionally, the studies on the pathogenic characteristics and contribution of pancreatic islet cells mainly focused on the abnormal biochemical reactions and physiological processes of such cell types in type 2 DM (Borg et al., 2001; Donath et al., 2003; Prentki and Nolan, 2006; Westermark and Westermark, 2008). According to these studies, there are four major pathogenic characteristics of pancreatic islet cells, including increased islet glucose metabolism (Forst et al., 2014), abnormal lipid signaling (Chakraborty et al., 2014), abnormal GLP-1 secretion (Trujillo and Nuffer, 2014), and compensatory feedback stimulation on parasympathetic and sympathetic neurons (Thorens, 2014). With the development of high-throughput sequencing technologies, more and more fundamental pathogenesis of type 2 DM at genetic and transcriptomic levels has been revealed. Apart from such transcription factors, genes regulating optimal glucose-responsive insulin secretion, like *IAPP*, *GLUT2*, *GAD65*, and *IA-2*, have also been identified to participate in T2D-associated pathogenesis (Clocquet et al., 2000). Therefore, the abnormal gene functions of pancreatic islet cells may be one of the major pathogenic factors for type 2 DM. However, as we all know, the cellular components of pancreatic islet cells are quite complicated involving various cell subtypes. Meanwhile, traditional studies all focused on the biological features (either at the cellular level or genetic level) of cell population, no matter pathogenic or not for individual cells. Therefore, these conventional studies may ignore some potential pathogenic factors and mistake non-pathogenic features due to normal or irrelevant cells' interferences.

Multiple previous studies have focused on single-cell analyses on pancreatic islets under physical or pathological conditions. With the development of single-cell techniques, the studies on pancreatic islets under either pathological or normal conditions have been extended to the single-cell level. Early in 2016, Segerstolpe et al. (2016) have identified some typical biomarkers to distinguish pancreatic islets under healthy and diabetic conditions. However, as limitations of this study, the authors only applied differential expression analyses and the t-SNE method to identify some potential biomarkers to reveal the heterogeneity (Segerstolpe et al., 2016). Apart from this study, further in 2017, another study extended to identify the specific biomarkers for T2D, confirming that genes are differentially expressed at the transcriptomics level not only between patients and controls but also among different cell types (Lawlor et al., 2017). In 2018, another single-cell gene expression analysis on T2D also tried to identify specific biomarkers for the prediction of cellular states of beta-cells, either healthy or T2D beta-cells (Ma and Zheng, 2018). The shortcomings of these two studies turn out to be a lack of quantitative standards establishment, making it still quite hard to predict T2D using single-cell transcriptomics data.

To overcome the limitations of previous studies mentioned earlier, in this study, for the first time, we used the single-cell sequencing results from one previous study (Xin et al., 2016) and tried to extend their analyses at two levels: 1) using multiple machine learning algorithms for deep analysis; 2) taking the pancreatic islets as a whole and did not distinguish different cell subtypes. We extended the classification and prediction of cellular states from just beta cells to multiple cell types, including human pancreatic alpha, beta, delta, and PP cells. Also, different from previous studies, we did not just focus on the pathogenic effects of T2D on beta cells but tried to reveal the general comprehensive pathogenic effects on all the cells from the pancreatic islets. Although most of the previous studies identified that pancreatic islet B cells are the major participants in the pathogenesis of T2D, other cells, including alpha, delta, and PP cells, are also either shown to be correlated with the pathogenesis of T2D or may act as potential biomarkers for T2D due to their typical changes during the pathogenesis. Therefore, it is not only innovative but effective to reveal the comprehensive effects of T2D on pancreatic islets and identify more valuable biomarkers for such disease.

All in all, to remove the interferences caused by conventional bulk sequencing and analysis, we have tried to identify potential pathogenic factors of T2D from the transcriptomic profiling

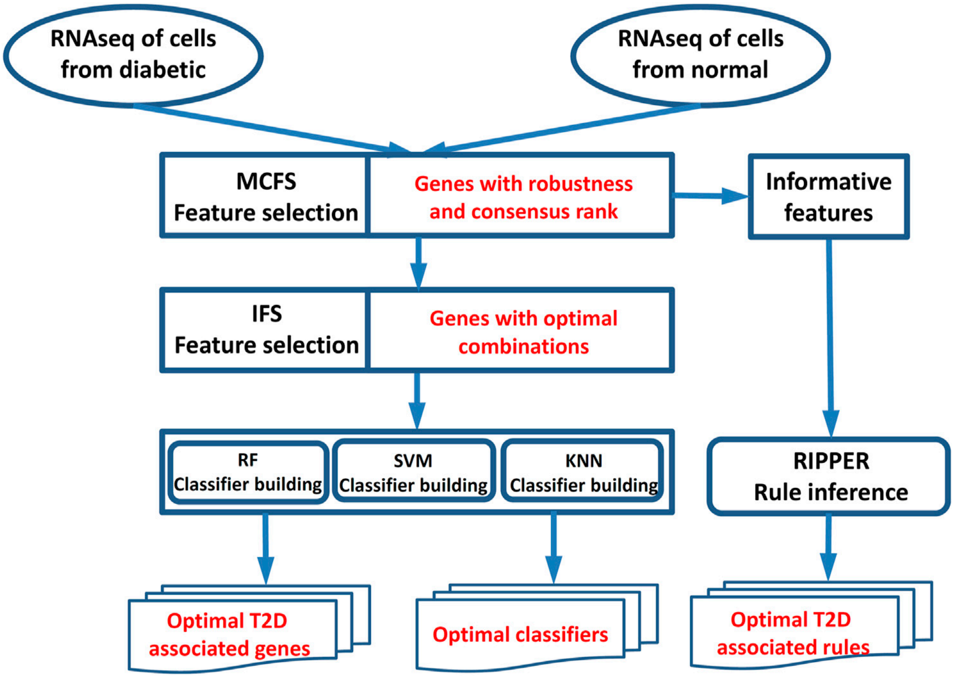

RNAseq of cells from diabetic
RNAseq of cells from normal
MCFS Feature selection
Genes with robustness and consensus rank
Informative features
IFS Feature selection
Genes with optimal combinations
RF Classifier building
SVM Classifier building
KNN Classifier building
RIPPER Rule inference
Optimal T2D associated genes
Optimal classifiers
Optimal T2D associated rules

FIGURE 1 | Workflow for key gene identification of type 2 diabetes. The MCFS method was used to evaluate the importance of all features (genes). On the one hand, the IFS method with SVM/RF/KNN was applied on the feature list yielded by the MCFS method to extract optimal T2D-associated genes and optimal classifiers. On the other hand, the informative features yielded by the MCFS method were fed into the Johnson reducer and RIPPER algorithms to construct optimal T2D-associated rules.

covering multiple cell subtypes at the single-cell level. Relied on single-cell RNA sequencing techniques and related public datasets (Xin et al., 2016), we investigated such datasets with several powerful machine learning algorithms. Different from previous studies, focusing on identifying biomarkers for distinguishing a tissue under normal or pathological conditions but not an entire tissue, which makes hard to detect biomarkers from a single-cell subtype in clinical applications, this study tried to identify the common transcriptomics characteristics across different cell types at the single-cell level for T2D. Biomarkers identified in this study may not be affected by the cell composition of the islet tissue that may vary among different individuals. In addition, our results revealed novel potential pathogenic mechanisms induced by newly recognized genes in a rule manner, which are always neglected by traditional bulk sequencing techniques. On the one hand, these results deepen our understanding on the etiology and pathogenesis of T2D. On the other hand, such identified new biomarkers can be potential candidates for further clinical application in the diagnosis of T2D using the transcriptomics information of the entire tissue, with no further cell separation and preprocessing required.

2 MATERIALS AND METHODS

In this study, we first used a feature selection method to analyze a RNA sequencing dataset of T2D for ranking the important genes associated with T2D, and these genes were further optimized for diabetes using incremental feature selection (IFS) (Liu and Setiono, 1998) with some supervised classifiers. In the end, we applied the rule learning method to generate interpretable classification rules for T2D. The whole process is illustrated in **Figure 1**.

2.1 Datasets

We downloaded the RNA sequencing data of 1,600 human pancreatic islet cells from the GEO (Transcript Expression Omnibus) database under the accession number of GSE81608 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81608 (Xin et al., 2016). There were 949 pancreatic islet cells from six T2D patients and 651 pancreatic islet cells from 12 non-diabetic donors. Within the 949 pancreatic islet cells from T2D patients, there were 569 alpha, 296 beta, 30 delta, and 54 PP cells. Within in the 651 pancreatic islet cells from non-diabetic donors, there were 377 alpha, 207 beta, 28 delta, and 39 PP cells. The expression levels of 39,851 genes were quantified as RPKM (Reads Per Kilo bases per Million reads). The processed gene expression profiles of these cells downloaded from https://ftp.ncbi.nlm.nih.gov/geo/series/GSE81nnn/GSE81608/suppl/GSE81608_human_islets_rpkm.txt.gz were used. Despite islet cells containing different cells, this work expects to identify the common gene signatures for T2D across multiple cell types.

2.2 Feature Selection

In this study, we first used the Monte Carlo feature selection (MCFS) (Draminski et al., 2008) to evaluate the importance of all

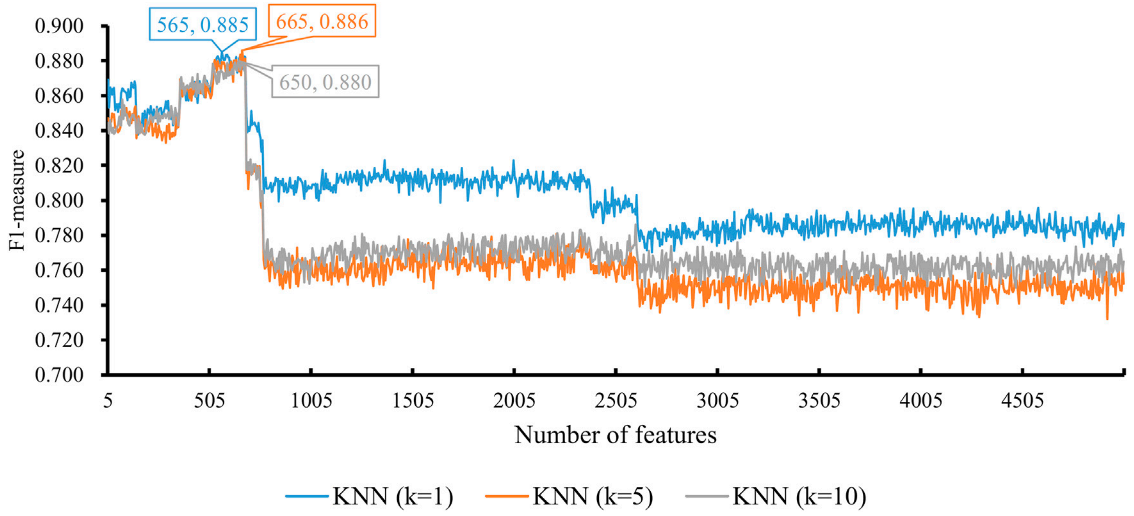


FIGURE 2 | Performance of KNN integrated in IFS using different numbers of features. The y-axis is F1-measure, and the x-axis is the number of participated features. *k* is the parameter of KNN, indicating the number of nearest neighbors that are used to make prediction. KNN can yield the best F1-measure of 0.886 when *k* = 5 and the top 665 features are used.

genes, obtaining a feature list and some informative genes expressed in diabetes. For the feature list, it was fed into the IFS (Liu and Setiono, 1998) with one classification algorithm to extract optimal genes that had a strong discriminate ability between diabetes and non-diabetes samples and construct an efficient classifier. On the other hand, repeated incremental pruning to produce error reduction (RIPPER) was employed to determine interpretable rules on gene expression patterns with informative features.

2.2.1 Monte Carlo Feature Selection

The investigated data contained 1,600 samples, each of which was represented by expression levels on lots of genes. Accordingly, the data can be summarized as a matrix with low row numbers and high column numbers. MCFS is deemed to be a powerful feature selection method to deal with such data. Thus, it was employed in this study. MCFS is a multivariate feature selection method based on bootstrap samples and decision trees, which focuses on selecting discriminate features for classification with robustness. In this feature selection algorithm, it generates multiple bootstrap sets, and on each bootstrap set, multiple decision trees are grown on smaller feature subsets randomly selected from original features. Then, the involvement of each feature in the decision trees shows a relative importance (RI) score, which indicates the overall number of splits involving this feature in all nodes of all constructed trees. The MCFS program was downloaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html. For convenience, default parameters were adopted.

The MCFS program was executed on the aforementioned RNA sequencing data. According to the output of the MCFS program, we can obtain the RI values of all features. Accordingly, features can be ranked in a list with the decreasing order of their RI values. Furthermore, it also provides the informative features, which are generated by a permutation test on class labels and one-sided Student's t-test. These features are always the top-ranking features in the list. We would adopt these features to construct classification rules *via* RIPPER.

2.2.2 Incremental Feature Selection

In this study, we performed IFS on the MCFS-generated feature list, denoted by $F = [f_1, f_2, \ldots, f_N]$ (N was the total number of features), to screen out a set of optimal features, which can accurately discriminate between diabetes and non-diabetes samples. Based on such list, we generated a series of feature subsets with step 5. Suppose there are m feature subsets $[F_1, F_2, \ldots, F_m]$, where the ith feature subset contains top $5 \times i$ features, that is, $F_i = [f_1, f_2, \ldots, f_{i \times 5}]$. Then, for a given classification algorithm, we built one classifier on samples represented by features from each feature subset and yielded the 10-fold cross-validation performance for evaluating this classifier. After all constructed feature subsets had been tested, the feature subset, on which the classifier provided the best performance, can be obtained. Such a feature subset was called the optimal feature subset for this classification algorithm, and the features inside were named as the optimal features. Furthermore, the classifier with the best performance was termed as the optimal classifier.

2.3 Classification Algorithm

For the IFS method, one classification algorithm was necessary. In this study, we tried three classic classification algorithms: 1) support vector machine (SVM) (Cortes and Vapnik, 1995), 2) K-nearest neighbor (KNN) (Cover and Hart, 1967), and 3) random forest (RF) (Breiman, 2001). Their brief descriptions were as follows.

2.3.1 Support Vector Machine

The SVM is a supervised learning model based on statistical learning theory and is widely used in many biological problems

(Pan and Shen, 2009; Mirza et al., 2015; Chen et al., 2017; Jia et al., 2018; Wei et al., 2018; Zhou et al., 2022a; Zhou et al., 2020b; Liu et al., 2021; Wang et al., 2021; Zhu et al., 2021; Li X. et al., 2022; Wu and Chen, 2022). Given a set of training samples, each training sample is assigned to positives or negatives. The SVM training algorithm fits a hyperplane that has the maximum margin between positives and negatives, where the generalization error becomes smaller when the margin is larger. The SVM generally is good at handling non-linear data, since it can first map the data in non-linear space to high-dimensional linear space by the kernel function and then fit a linear model in the high-dimensional space.

2.3.2 K-Nearest Neighbor

KNN is one of the simplest schemes for classifying samples. However, in many cases, it still can yield good performance. Given a training dataset, KNN directly uses samples in it to make prediction for any query sample, that is, KNN does not contain a learning procedure. Generally, it finds k training samples, which have the nearest distances (e.g., Euclidean distance) to the query sample. By counting the classes of these k training samples, the class with most votes is assigned to the query sample.

2.3.3 Random Forest

RF is another classic classification algorithm. In fact, it is an integrated algorithm, consisting of several decision trees. For constructing each decision tree, it randomly picks up samples from the training dataset, with replacement, to constitute the basic dataset. The tree is extended at each node by selecting an optimal split on one feature among the randomly selected features. RF integrates the predictions of all decision trees with majority voting. RF is deemed as a powerful classification algorithm and has wide applications in tackling many biological problems (Kandaswamy et al., 2011; Casanova et al., 2014; Marques et al., 2016; Jia et al., 2020; Liang et al., 2020; Zhang et al., 2021b; Chen et al., 2021; Onesime et al., 2021; Chen et al., 2022; Ding et al., 2022; Yang and Chen, 2022).

To quickly implement the aforementioned three classification algorithms, we employed the corresponding packages in scikit-learn (https://scikit-learn.org/stable/). Some main parameters were tuned for extracting optimal parameters.

2.4 Johnson Reducer and Repeated Incremental Pruning to Produce Error Reduction Algorithms

Classification algorithms mentioned in **Section 2.3** are powerful to construct efficient classifiers. However, we cannot understand their principles because they are black-box algorithms. In this case, few clues for uncovering essential differences between T2D patients and non-diabetic donors can be obtained. In view of this, we further adopted rule learning algorithms to investigate the RNA sequencing data. Although it is generally weaker than the aforementioned algorithms, it can provide rules that clearly indicate special expression patterns on T2D patients, thereby improving our understanding on T2D. The procedures were described in the following sections.

As mentioned in Section 3.2.1, the MCFS method can select some informative features. These features are quite essential to describe the characteristics of the dataset. Here, we used these features to construct classification rules *via* RIPPER algorithm (Cohen, 1995). Before that, the Johnson reducer algorithm (Johnson, 1974) was applied on the informative features to select the most important features, which had the similar classification ability compared to the original informative features. The selected features were fed into the RIPPER algorithm. RIPPER, proposed by Cohen (1995), is a rule learning algorithm which is capable of handling large noisy datasets effectively. RIPPER is the improved version of IREP (Johannes and Widmer, 1994) which combines both the separate-and-conquer technique used first in the relational learner FOIL (Quinlan, 1990) and the reduced error pruning strategy proposed by Brunk and Pazzani (1991). In RIPPER, the training set is first split into growing and pruning sets. Then, repeat the rule grow phase and rule prune phase until no positive samples are left in the growing set, or the description length (DL) is 64 bits greater than the smallest DL found so far, or the error rate is greater than 50%. In the rule grow phase, one rule is generated by greedily adding conditions to the rule that achieves the highest FOIL's information gain. In the rule prune phase, the rule is pruned using reduced error pruning. Finally, global optimization strategy is applied to further prune the rule set. The aforementioned procedures for constructing rules are also implemented in the MCFS program, that is, the set of rules is one output of the MCFS program.

2.5 Performance Measurement

In this study, we used six measurements to evaluate the performance of all classifiers under 10-fold cross-validation (Kohavi, 1995; Li Z. et al., 2022; Ding et al., 2022; Tang and Chen, 2022), including sensitivity (SN) (same as recall), specificity (SP), accuracy (ACC), Matthew correlation coefficient (MCC), precision, and F1-measure (Matthews, 1975; Zhao et al., 2018; Zhao et al., 2019; Jia et al., 2020; Liang et al., 2020; Zhang et al., 2021a; Zhang et al., 2021c; Pan et al., 2021). Their formulations are written as follows:

$$SN = Recall = \frac{TP}{TP + FN}, \tag{1}$$

$$SP = \frac{TN}{TN + FP}, \tag{2}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

$$F1 - measure = \frac{2 \times Recall \times Precision}{Recall + Precision}, \tag{6}$$

where TP represents the number of truly positive samples, FP represents the number of false-positive samples, TN represents the number of truly negative samples, and FN represents the number of false-negative samples. Among these six measurements, we selected F1-measure as the key one, whereas others were provided for reference.

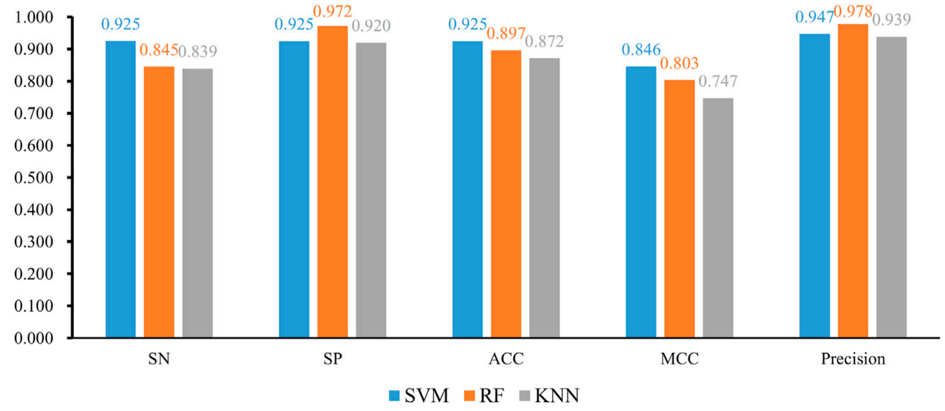


FIGURE 3 | Bar chart to show five measurements of three optimal classifiers based on different classification algorithms.

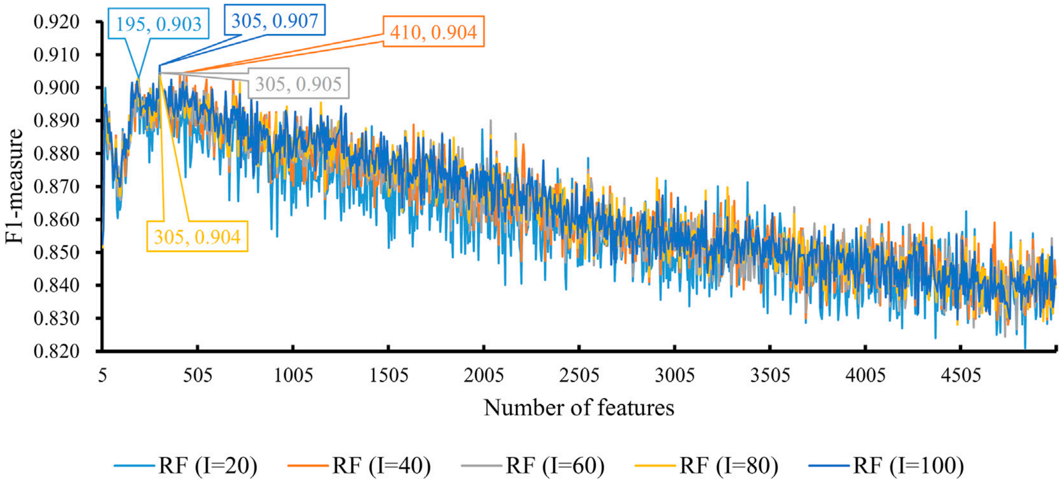


FIGURE 4 | Performance of RF integrated in IFS using different numbers of features. The y-axis is F1-measure, and the x-axis is the number of participated features. *I* is the parameter of RF, indicating the number of decision trees. RF can yield the best F1-measure of 0.907 when *I* = 100 and the top 305 features are used.

2.6 Gene Ontology Enrichment Analysis on Optimal Genes

Some rules can be extracted *via* the Johnson reducer and RIPPER algorithms, which involved several features (genes), called rule genes, in the following text. We performed Gene Ontology (GO) enrichment analysis using R package *topGO* (http://bioconductor.org/packages/release/bioc/html/topGO.html, v.2.24.0) on these rule genes. The genes of interest were set as rule genes, and the gene background was set as all the available genes. The *p*-value threshold was set at 0.001.

3 RESULTS

T2D is one type of DM and makes up most DM cases. In this study, we investigated potential pathogenic factors of T2D at the single-cell level by analyzing a single-cell RNA sequencing dataset. Such dataset contained 1,600 single cells, including 949 cells from T2D patients and 651 cells from normal controls. It was analyzed by some powerful machine learning algorithms, including MCFS (Draminski et al., 2008), SVM (Cortes and Vapnik, 1995), KNN (Cover and Hart, 1967), RF (Breiman, 2001), and RIPPER (Cohen, 1995). The entire procedure is shown in **Figure 1**. On one hand, we obtained some T2D-associated genes, which can be novel biomarkers of T2D. On the other hand, some interesting rules were constructed, which can uncover different expression patterns in T2D patients and normal controls. This section gives the detailed results of these procedures.

3.1 Results of the Monte Carlo Feature Selection Method

The MCFS method was directly applied to the RNA sequencing data to analyze the importance of all features (genes). Each gene was assigned a RI score. A total of 26,978 genes were assigned RI scores larger than zero. These genes and their RI scores are

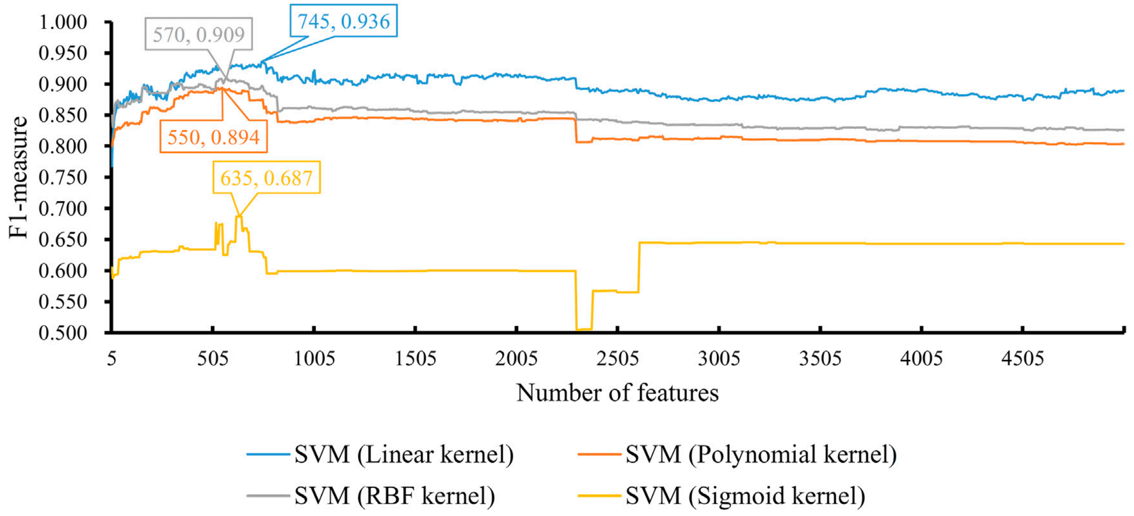


FIGURE 5 | Performance of SVM integrated in IFS using different numbers of features. The y-axis is F1-measure, and the x-axis is the number of participated features. SVM can yield the best F1-measure of 0.936 when the kernel is a linear function and the top 745 features are used.

provided in **Supplementary Table S1**. Because the RI scores of the rest genes were zero, meaning their associations for the identification of T2D samples were very weak, they were discarded. A feature list was generated by sorting the remaining 26,978 genes in the decreasing order of their RI scores, which is also provided in **Supplementary Table S1**.

In addition to the feature list, the MCFS method can output some informative features. For investigating RNA sequencing data, 235 informative features were extracted by the MCFS method, which were the top 235 genes listed in **Supplementary Table S1**.

3.2 Results of the Incremental Feature Selection Method

To further extract optimal features, the IFS method combined with one classification algorithm was employed. Here, we tried three classification algorithms: SVM, KNN, and RF. Some main parameters of each algorithm were tuned. In detail, for SVM, four kernels were attempted, including linear, polynomial, RBF, and sigmoid kernels. The parameter *k* for KNN was set to 1, 5, and 10, and the parameter, number of decision trees (I), for RF was set to 20, 40, 60, 80, and 100. Because the feature list contained a huge number of features, we only considered the top 5,000 features in this study to save time. Several feature subsets were constructed using step 5.

When the classification algorithm was KNN, several KNN classifiers with a certain parameter *k* were constructed on all feature subsets. All these classifiers were evaluated by 10-fold cross-validation. The obtained six measurements are listed in **Supplementary Table S2**. For an easy observation, we plot a curve for KNN with a certain parameter *k*, as shown in **Figure 2**, in which the F1-measure was set to the y-axis and the number of features was set to the x-axis. We can see that when *k* = 1, 5, and 10, the highest F1-measure was 0.885, 0.886, and 0.880, respectively. Thus, the KNN classifier with *k* = 5 provided the best performance. Such classifier used the top 665 features (genes) in the feature list. These features were the optimal features for KNN. The other five measurements are illustrated in **Figure 3**. Except MCC, all measurements exceeded 0.8, implying the good performance of such KNN classifiers. Furthermore, it can be observed from **Figure 2** that the IFS curves of KNN with different parameters *k* had a common feature. The curve followed a sharp decreasing trend before about top 600 features were used. The top features in the list were highly related to class labels (T2D patients and non-diabetic patients in this study), and a simple scheme based on these features, as KNN used, can correctly predict the cells of T2D patients and non-diabetic patients. However, when features with low ranks, which had low relevance to class labels, were added, KNN cannot exclude interference information contained in these features as KNN has no training procedures, inducing the quick descent of its performance. In this study, the set containing about top 600 features was a pivotal point for KNN. After this point, the performance of KNN followed a sharp decreasing trend.

We also tried another classification algorithm, RF. The same IFS procedure was conducted on this algorithm. The obtained measurements are listed in **Supplementary Table S3**. Likewise, a curve was plotted for RF with a certain number of decision trees, as shown in **Figure 4**. It can be observed that when *I* = 20, 40, 60, 80, and 100, the highest F1-measure was 0.903, 0.904, 0.905, 0.904, and 0.907. The RF classifier with *I* = 100 provided the highest performance. The top 305 features in the list were adopted in this classifier and were termed as optimal features for RF. Evidently, such an RF classifier was superior to the best KNN classifiers mentioned earlier. Furthermore, the other five measurements of this RF classifier are shown in **Figure 3**. All measurements were higher than 0.8, suggesting the better performance of this classifier than the aforementioned KNN classifier.

TABLE 1 | Top seven genes among the optimal genes for SVM.

Rank	Gene ID	Gene symbol	RI
1	100128906	LOC100128906	0.1140
2	100873254	MTND4P24	0.1046
3	100271063	RPS14P1	0.1032
4	100652939	MTND2P28	0.0979
5	285045	LINC00486	0.0959
6	729898	ZBTB8OSP2	0.0954
7	391524	THRAP3P1	0.0862

Finally, we conducted the same IFS procedure for SVM. The measurements are listed in **Supplementary Table S4**. Similarly, for each SVM with a certain kernel, a curve was plotted, as shown in **Figure 5**. With four different kernels, SVM yielded the highest F1-measure of 0.936, 0.894, 0.909, and 0.687. The SVM with a linear kernel provided the best performance. Also, such performances were based on the top 745 features in the list. Accordingly, they were called the optimal features for SVM. Furthermore, the performance of this SVM classifier was better than that of the aforementioned KNN and RF classifiers. The same conclusion can be obtained according to the five measurements of such SVM classifiers, illustrated in **Figure 3**. Due to the best performance of the SVM with its optimal 745 genes, these genes were quite important for investigating T2D at the single-cell level. The top seven genes are listed in **Table 1**.

With the earlier IFS results with different classification algorithms using various parameters, the SVM with linear kernel and top 745 features provided the best performance of F1-measure 0.936. The ACC and MCC of such classifier were 0.925 and 0.846, respectively. Other three measurements, SN, SP, and precision were 0.925, 0.925, and 0.947, respectively. These measurements suggested the excellent performance of this classifier, and it can be an efficient tool to identify cells of T2D patients.

3.3 Classification Rules

Although we can construct efficient classifiers to identify cells of T2D patients through three classification algorithms, these classifiers were absolute black-box algorithms, which prevented us from uncovering the essential differences between cells of T2D patients and non-diabetic donors. As mentioned in **Section 2.4**, rule learning algorithms were employed.

According to the output of the MCFS program, 235 features were selected as informative features. To test the utility of the classification rules yielded by Johnson reducer and RIPPER algorithms, we performed the 10-fold cross-validations three times, obtaining the F1-measure of 0.910, which was lower than that of the optimal SVM classifier but higher than that of the optimal KNN and RF classifiers. The SN was 0.898, SP was 0.891, ACC was 0.895, MCC was 0.784, and precision was 0.923. Although such performance was lower than that of the optimal SVM classifier, the RIPPER algorithm can construct a group of rules, which made the classification procedure completely open and provided more insights. Thus, the Johnson reducer and RIPPER algorithms were applied to all samples, producing nine different classification rules, as listed in **Table 2**. These rules are able to accurately screen patients with T2D from non-diabetic population. Although these rules were mainly for non-diabetes, based on the aforementioned evaluation results (SP = 0.891), it was believed that these rules were statistically shown to cover almost all possible non-diabetes samples. Thus, investigation on these rules can also figure out the characteristics of T2D patients in an opposite aspect.

TABLE 2 | Nine classification rules for diabetes generated by the RIPPER algorithm.

Rule	Criteria	Patient
Rule 1	Gene Id 100128906 (LOC100128906) ≥ 2.7722 Gene Id 326307 (RPL3P4) ≤ 15.2306 Gene Id 8781 (PSPHP1) ≥ 0.0965 Gene Id 100873065 (PTCHD1-AS) ≤ 0.1036	Non-diabetes
Rule 2	Gene Id 100462954 (MICOS10P3) ≥ 2.0984 Gene Id 1487 (CTBP1) ≤ 17.3460 Gene Id 326307 (RPL3P4) ≤ 6.2868 Gene Id 100873254 (MTND4P24) ≥ 3.0364	Non-diabetes
Rule 3	Gene Id 100128906 (LOC100128906) ≥ 49.6340 Gene Id 143244 (EIF5AL1) ≥ 1.0987 Gene Id 486 (FXYD2) ≤ 152.8666 Gene Id 326307 (RPL3P4) ≤ 11.3894 Gene Id 6126 (RPL9P7) ≤ 103.5050	Non-diabetes
Rule 4	Gene Id 100128906 (LOC100128906) ≥ 3.0256 Gene Id 326307 (RPL3P4) ≤ 22.4381 Gene Id 100128906 (LOC100128906) ≥ 225.8732 Gene Id 388147 (RPL9P9) ≤ 50.3934 Gene Id 100271332 (RPL36AP21) ≥ 1.7952 Gene Id 222901 (RPL23P8) ≤ 2.6067	Non-diabetes
Rule 5	Gene Id 100652939 (MTND2P28) ≥ 450.8125 Gene Id 4574 (MT-TS1) ≤ 445.4115 Gene Id 1487 (CTBP1) ≤ 37.6438	Non-diabetes
Rule 6	Gene Id 285045 (LINC00486) ≤ 0.0930 Gene Id 100873254 (MTND4P24) ≤ 28.2479 Gene Id 653147 (RPL26P30) ≥ 5.1856 Gene Id 285900 (RPL6P20) ≥ 0.4760 Gene Id 643932 (RPS3AP20) ≥ 5.5063	Non-diabetes
Rule 7	Gene Id 100128906 (LOC100128906) ≥ 3.0256 Gene Id 440737 (RPL35P1) ≥ 4.118 Gene Id 100271003 (RPL34P18) ≥ 9.0166	Non-diabetes
Rule 8	Gene Id 100128906 (LOC100128906) ≥ 109.2232 Gene Id 100873254 (MTND4P24) ≤ 28.3353 Gene Id 644972 (RPS3AP26) ≥ 53.5552 Gene Id 644604 (EEF1A1P12) ≤ 7.9556	Non-diabetes
Rule 9	Others	Diabetes

3.4 Comparison of Classifiers With Informative Features

The MCFS method can directly output some informative features. These features can capture essential information of the dataset. Here, as mentioned in **Section 3.3**, 235 features were selected as informative features. We can directly use them to construct

TABLE 3 | Performance of classifiers using informative features yielded by the MCFS method.

Classification algorithm	F1-measure	Decrement[a]
KNN (k = 1)	0.849	0.036
KNN (k = 5)	0.839	0.047
KNN (k = 10)	0.847	0.033
RF (l = 20)	0.889	0.014
RF (l = 40)	0.891	0.013
RF (l = 60)	0.894	0.011
RF (l = 80)	0.894	0.010
RF (l = 100)	0.897	0.010
SVM (linear kernel)	0.882	0.054
SVM (polynomial kernel)	0.859	0.035
SVM (RBF kernel)	0.886	0.023
SVM (sigmoid kernel)	0.631	0.056

[a]Numbers listed in this column indicate the difference of F1-measure yielded by the optimal classifier and that listed in the second column of this table.

classifiers with different classification algorithms. These classifiers were also evaluated by 10-fold cross-validation. The main measurement, F1-measure, of these classifiers is listed in **Table 3**. For KNN, F1-measure varied between 0.839 and 0.849. The F1-measure of RF changed between 0.889 and 0.897. Also, SVM provided the F1-measure varying between 0.631 and 0.886. Compared with the F1-measure yielded by the optimal classifier based on the corresponding classification algorithm, the classifier using informative features generated a lower F1-measure, suggesting that such a classifier was inferior to the optimal classifier. The employment of the IFS method can help us construct more efficient classifiers.

4 DISCUSSION

As we have described earlier, we applied our newly presented computational framework to the expression profiling data of more than 1,600 single pancreatic islet cells, constituting 949 diabetic cells and 651 non-diabetic cells (Xin et al., 2016). Based on such a bioinformatics approach, we not only screened out a group of discriminative genes that have distinctive expression patterns in diabetic or non-diabetic cells but also set up a series of quantitative rules for the recognition of pathogenic cells at the single-cell level. According to recent literature reports, several identified genes and established rules could be validated by existing experimental datasets, indicating the efficacy and accuracy of our analysis. The detailed functional analysis and evaluation of each predicted genes with high informative rank and their optimal rules in the expression pattern have been summarized and introduced in the following sections.

4.1 Analysis of Optimal Type 2 Diabetes-Associated Genes

Because the optimal SVM classifier provided the best performance, which used top 745 features (genes), we focused on these 745 genes. However, it is impossible to analyze them one by one. Here, only top seven genes were analyzed, which are listed in **Table 1**.

The first predicted gene, *WDR45-like pseudogene* (100128906), is the pseudogene of gene *WDR45*. According to recent publications, it encodes a functional lncRNA associated with the regulation of *WDR45* (Tsuyuki et al., 2014; Lebovitz et al., 2015). *WDR45* has been functionally related to autophagy (Lebovitz et al., 2015). Considering that abnormal autophagy has been well known to contribute to the pathogenesis of T2D (Lee, 2014), it is reasonable to speculate that the expression level of *WDR45* and its upstream regulator (i.e., our predicted gene *LOC100128906*) may have quite different expressions in diabetic pancreatic islets cells compared to normal cells.

The next identified gene is *MTND4P24* (100873254), which is shown to have quite different expression levels in diabetic and normal tissues containing multiple cell subtypes. As an lncRNA-encoding pseudogene, the expression level of such a gene is able to reflect the regulatory ability of lncRNAs on its target gene, *MT-ND4* (Torrell et al., 2013; Mella et al., 2016). Recent publications also confirmed that the expression level of the target gene *MT-ND4* is functionally related to cellular insulin sensitivity in rat models (Houstek et al., 2012). Therefore, as one regulator of *MT-ND4*'s expression, the expression pattern of *MTND4P24* may involve in the pathogenic insulin sensitivity decreasing in type 2 diabetic cells. Similarly, a homolog of *MTND4P24* and *MTND2P28* (100652939) has also been predicted to have different expression levels in multiple cell subtypes from pathogenic or normal pancreatic islets. Considering its similar regulatory mechanisms and the biological function of MTND2, it is also quite convincing to regard such a gene as a potential distinctive standard for diabetic and non-diabetic cells (Mathews et al., 2005).

The predicted gene, *RPS14P1* (100271063), is also a pseudogene, contributing to the regulation of ribosomal protein S14's expression (Aubert et al., 1992). Meanwhile, the function of ribosomal protein S14 is widely reported to participate in p53-dependent cell-cycle arrest by interacting with *MDM2* (Zhou et al., 2013), which is abnormally activated during the pathogenesis of diabetes (Golubnitschaja et al., 2006; Garufi et al., 2017). Thus, it is a reasonable assumption that ribosomal protein S14 together with *RPS14P1* has different expression levels in normal and diabetic cells.

Apart from such predicted pseudogenes, we also identified some functional lncRNAs that may have different expression patterns in normal and diabetic cells. LINC00486 (285045) is a predicted lncRNA that contributes to the distinction of normal and diabetic cells. According to recent publications, various functional lncRNAs (Liu et al., 2014; Pullen and Rutter, 2014), including LINC00486, have been confirmed to contribute to the initiation and progression of T2D (Pullen and Rutter, 2014).

The following predicted gene, named *ZBTB8OSP2* (729898), is a pseudogene and has been reported to contribute to anti-saccade response and eating disorders (Cornelis et al., 2014; Broer and van Duijn, 2015). As a transcriptional regulator for *ZBTB8*, such genes may indirectly contribute to a specific complication of T2D, the refractory diabetes insipidus, especially in adolescent male patients (Soto et al., 2014). Therefore, we can infer that such genes together with their downstream binding targets may have respective specific expression patterns in normal and diabetic cells.

The next predicted gene is *THRAP3P1* (391524), the pseudogene of *THRAP3*. The post-transcriptional regulatory target of *THRAP3* has been confirmed to dock on phosphoserine 273 of PPAR-gamma and further contribute to the pathogenic programming of diabetic genes, inducing insulin resistance (Choi et al., 2014). Therefore, to accomplish the regulatory role, such gene has a high expression level in normal cells compared to diabetic cells.

4.2 Specific Role of Pseudogenes in Type 2 Diabetes-Associated Genes

As we have discussed earlier, we identified multiple pseudogenes associated with T2D. Pseudogenes are nonfunctional segments with similar or reverse sequences of actual coding genes. The biological functions of pseudogenes are still unclear. It has only been speculated that pseudogenes participate in the post-transcriptional regulation *via* generating siRNAs, piRNAs, microRNAs, or other small RNAs (Guo et al., 2009). Although pseudogenes cannot generate protein products, the regulatory effects of such group of genes may still be significant under physical and pathological conditions (Tay et al., 2014). For transcriptomics analyses, especially for single-cell transcriptomics analyses, multiple pseudogenes have been identified as candidate biomarkers for different systematic diseases in studies specifically focusing on pseudogenes' effects (Kalyana-Sundaram et al., 2012; Poliseno et al., 2015). For most previous studies, the pseudogenes were removed in the data preprocessing. Therefore, most previous studies have not identified a lot of pseudogenes as potential candidate biomarkers for diabetes. In our study, we did not filter out the pseudogenes and for the first time confirmed that pseudogenes with potential transcriptomic regulatory effects may further contribute to the regulation of specific diseases *via* regulating the biological functions of their respective recognized protein-coding genes.

4.3 Comparison With Previously Reported Type 2 Diabetes Biomarkers

Here, in this study from other perspective of view, we applied several machine learning algorithms to identify new potential biomarkers for T2D patients. Multiple previous publications have already identified a group of T2D biomarkers such as HbA1c, advanced glycation end-products (AGEs), and pigment epithelial-derived factor (PEDF) (Lyons and Basu, 2012). Also, for the publication from which we retrieved the single-cell sequencing data, unique biomarkers like *LINC00486*, *ZNF445*, and *SYBU* have also been identified for T2D (Xin et al., 2016). Compared with these prediction results, first, we identified a group of confirmed biomarkers like *LINC00486*, validating the efficacy and accuracy of our results. Second, we identified a group of new biomarkers like *MTND4P24* and *THRAP3P1*. Although such genes have been shown to be functionally correlated with T2D, previous studies have not identified such genes as potential biomarkers of T2D. There are two major advantages of our studies compared to previous studies, which may lead us to find novel biomarkers:

1) First, compared with previous studies, we used the single-cell level data with the gene expression profiling of different cells and not just an averaged comprehensive value for each patient. Therefore, we can identify potential biomarkers that are missing due to the averaging procedures.
2) Second, due to the sample size and cell type distribution, it is not proper to use feature selection and machine learning models for distinguishing each cell type independently. An integration of all the cell types may lead to a more reasonable result with effective biomarkers with clinical application potentials.

Such advantages explained why we identified novel protein biomarkers to distinguish T2D patients from normal controls. As we have discussed earlier, some identified biomarkers have been functionally correlated with T2D, implying that it is reasonable to regard such genes/transcripts as potential biomarkers for T2D.

4.4 Analysis of Optimal Type 2 Diabetes-Associated Rules

We also screened out a group of functional quantitative rules of the gene expression pattern to distinguish non-diabetic cells from diabetic ones with more interpretability, which are listed in **Table 2**. Many qualitative rules can be validated according to the gene expression level in existing databases and recent reports on gene expression trends, which support the efficacy and accuracy of the rules. The detailed analysis of each expression rule is widely discussed as follows:

The first rule (rule1) involved four genes including *LOC100128906* [(100128906), *RPL3P4* (326307), *PSPHP1* 8781], and *PTCHD1* (100873065). As mentioned earlier, gene *LOC100128906* has been reported to have quite different transcriptomics patterns between normal and diabetic cells, inhibiting autophagy (Lebovitz et al., 2015). As the antagonistic gene of diabetes-associated autophagy, such genes are reasonable to have high expression in normal cells compared to diabetic cells. As for gene *RPL3P4*, the regulatory target of such pseudogene, *RPL3* has been reported to have a quite low expression level in diabetic cells compared to normal cells (Tsai et al., 1994), corresponding with this rule. As for *PSPHP1* (8781), it has been shown to be associated with the macrophage-related inflammation processes (Walker et al., 2015). Considering that during the initiation and progression of diabetes, regional and systematic inflammation have been widely observed (Donath et al., 2003; Lontchi-Yimagou et al., 2013), it is reasonable to predict such genes as quantitative parameters for the distinction of non-diabetes and diabetes. As for *PTCHD1*, although no direct evidence confirms its contribution on diabetes, it has been confirmed that such a

TABLE 4 | Significant Gene Ontology enrichment analysis result on rule genes.

GO ID	Term	*p*-value	Cluster
GO:1903408	Positive regulation of sodium: potassium-exchanging ATPase activity	5.30E-04	BP
GO:0045901	Positive regulation of translational elongation	7.00E-04	BP
GO:0045905	Positive regulation of translational termination	7.00E-04	BP

gene is associated with the eye and ear complications of diabetes (Gambin et al., 2017), consistent with this rule.

As for the second rule (rule2), four genes were involved including *MICOS10P3* (100462954), *CTBP1* (1487), *RPL3P4* (326307), and *MTND4P24* (100873254). Few publications have reported the biological contribution of *MICOS10P3*; therefore, it is hard to interpret such gene's contribution on T2D. As for gene *CTBP1*, it has been reported to participate in the abnormal phosphorylation processes (Kim et al., 2013) and shows a quite high expression level in diabetic cells compared to normal controls. As for gene *RPL3P4*, the regulatory target of such a pseudogene, *RPL3* has been reported to have a quite low expression level in diabetic cells compared to normal cells (Tsai et al., 1994), corresponding with such a rule. *MTND4P24* and its homolog, *MTND5P11*, have been confirmed to regulate a group of functional mitochondrial-encoded NADH ubiquinone oxidoreductase. According to recent publications, during the pathogenesis of diabetes, *MT-ND4* has a quite low-expression pattern and on the contrary, *MT-ND5* has a relevantly higher expression level, corresponding with the prediction expression level of their agonists individually (Elango et al., 2014; Urbanova et al., 2017).

In the third rule (rule3), apart from genes we have discussed earlier, the gene *EIF5AL1* (143244) has also been predicted to have a higher expression pattern in normal cells but not in diabetic cells. Considering the abnormal endocrine stress responses of diabetic cells (Siddiqui et al., 2015), the lower expression level of *EIF5AL1* may also contribute to the identification of diabetic cells. *FXYD2* (486) has been shown to contribute to the pathogenesis of diabetes (Ding et al., 2019). Another specific gene in rule3 is the homolog of *RPL3P4*, *RPL9P7*, which may also participate in the regulation of the pathogenesis of T2D with similar expression patterns to *RPL3P4*.

From the fourth to eighth rules, most of the involved genes occurred in the top three rules or were the top T2D-associated genes just with different combination patterns. Specific genes, like *RPL9P9* (388147) and *RPL36AP21* (100271332) for rule4, *MT-TS1* (4574) for rule5, *RPL26P30* (653147) and *RPL6P20* (285900) for rule6, *RPL35P1* (440737) for rule7, and *RPS3AP26* (644972) for rule8, have been identified in our quantitative rules. As we can see from such typical rule associating biomarkers, most of the genes are ribosome-associated genes like *RPL3P4* (326307) as discussed earlier. Although no direct evidence confirmed the associations between such genes and T2D, it is still reasonable to speculate that such genes may play an irreplaceable role in the identification of T2D. As for *MT-TS1*, such genes have already been reported as potential biomarkers for T2D (Mannino and Sesti, 2012), corresponding with our prediction.

4.5 Potential Applications of Identified Type 2 Diabetes-Associated Genes and Rules

There are two potential applications for identified T2D-associated genes: 1) potential biomarkers for T2D diagnosis and monitoring; 2) potential drug target for T2D therapy.

For the identified T2D-associated genes, considering that such genes are identified from pancreatic tissues, they can reflect the original tissue alterations during T2D initiation and progression. Therefore, such genes can be used as biomarkers for direct pancreatic islet biopsy examinations. Apart from that, the candidate genes as potential drug targets can also be manually regulated to prevent the initiation and progression of T2D. Using high-throughput drug screening, antibodies or chemicals specifically targeting the candidate genes can be identified and developed as potential target drugs for T2D.

For the quantitative T2D-associated rules, although we have already identified a group of genes associated with T2D, it is still quite difficult to diagnose T2D. With specific quantitative rules, the identification of T2D patients can be more accurate and efficient. Also, the rules can also be summarized as clinical guidelines for T2D diagnosis using pancreatic tissue single-cell sequencing techniques.

4.6 Functional Interpretation of Significant Rule Genes

As listed in **Table 2**, we identified quantitative rules associated with T2D. The GO enrichment analyses on rule genes were conducted. **Table 4** lists the enriched GO terms of these rule genes. It was indicated that most rules are shown to be associated with ribosome-associated biological processes. According to recent publications, ribosome-associated biological processes have been widely shown to be associated with the pathogenesis of T2D. In 2019, in a metabolic study on pancreatic tissues, ribosome-associated genes have been shown to participate in the ERK/hnRNPK/DDX3X pathway in pancreatic islet cells and further regulated the initiation and progression of T2D (Good et al., 2019), consistent with our results. Apart from that, in 2020, DIMT1, as a regulator of ribosomal biogenesis has been shown to participate in the physical biological processes of pancreatic tissue, further validating our results.

4.7 Limitations of Current Analyses

In this study, for the first time, we adopted several machine learning algorithms to identify disease-specific biomarkers using the mixed single-cell sequencing data. Such analyses may not only identify biomarkers from the single-cell level, getting rid of the bias generated by the averaged transcriptomics using the bulk sequencing method, but also overcome the sample size restriction of traditional single-cell analysis. Compared with traditional single-cell analysis, we did not focus on the classification of different cell subtypes but just the patients and control subjects, improving the analysis accuracy. However, there still remain three major limitations of current analyses on pancreatic single-cell sequencing data:

1) First, the dataset we used is still a relatively small dataset, with around 20 subjects. A larger single-cell sequencing dataset may improve the efficacy and accuracy of our results.
2) Second, the number of cells in each group is not balanced in the raw dataset. Although in the original publications the authors have claimed that the sampling procedure does not affect the distribution of cell subgroups in each subject, a more balanced dataset may perform better.
3) Single-cell sequencing always misses a lot of genes at low-expression levels which cannot be detected at the single-cell level but can be identified in bulk sequencing. Our analyses may also lose the gene expression profiling and analysis on such low-expression genes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81608.

AUTHOR CONTRIBUTIONS

Y-DC designed the study. ZL and XP performed the experiments. ZL analyzed the results. ZL and XP wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA26040304 and XDB38050200), the National Key R&D Program of China (2018YFC0910403), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.890901/full#supplementary-material

REFERENCES

American Diabetes Association (2014). Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* 37 (Suppl. 1), S81–S90. doi:10.2337/dc14-S081

Andersen, M. K., Pedersen, C.-E. T., Moltke, I., Hansen, T., Albrechtsen, A., and Grarup, N. (2016). Genetics of Type 2 Diabetes: the Power of Isolated Populations. *Curr. Diab Rep.* 16, 65. doi:10.1007/s11892-016-0757-z

Aubert, D., Bisanz-Seyer, C., and Herzog, M. (1992). Mitochondrial Rps14 Is a Transcribed and Edited Pseudogene in *Arabidopsis thaliana*. *Plant Mol. Biol.* 20, 1169–1174. doi:10.1007/bf00028903

Boden, G. (1997). Role of Fatty Acids in the Pathogenesis of Insulin Resistance and NIDDM. *Diabetes* 46, 3–10. doi:10.2337/diabetes.46.1.3

Borg, H., Gottsäter, A., Landin-Olsson, M., Fernlund, P., and Sundkvist, G. (2001). High Levels of Antigen-specific Islet Antibodies Predict Futureβ -Cell Failure in Patients with Onset of Diabetes in Adult Age1. *J. Clin. Endocrinol. Metabolism* 86, 3032–3038. doi:10.1210/jcem.86.7.7658

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Broer, L., and Van Duijn, C. M. (2015). GWAS and Meta-Analysis in Aging/Longevity. *Adv. Exp. Med. Biol.* 847, 107–125. doi:10.1007/978-1-4939-2404-2_5

Brunk, C. A., and Pazzani, M. J. (1991). "An Investigation of Noise-Tolerant Relational Concept Learning Algorithms," in Proceedings of the Eighth International Conference, Evanston, Illinois, June, 1991, 389–393. doi:10.1016/b978-1-55860-200-7.50080-5

Casanova, R., Saldana, S., Chew, E. Y., Danis, R. P., Greven, C. M., and Ambrosius, W. T. (2014). Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses. *PLoS One* 9, e98587. doi:10.1371/journal.pone.0098587

Chakraborty, C., Doss, C. G. P., Bandyopadhyay, S., and Agoramoorthy, G. (2014). Influence of miRNA in Insulin Signaling Pathway and Insulin Resistance: Micro-molecules with a Major Role in Type-2 Diabetes. *WIREs RNA* 5, 697–712. doi:10.1002/wrna.1240

Chen, L., Li, Z., Zhang, S., Zhang, Y. H., Huang, T., and Cai, Y. D. (2022). Predicting RNA 5-methylcytosine Sites by Using Essential Sequence Features and Distributions. *Biomed. Res. Int.* 2022, 4035462. doi:10.1155/2022/4035462

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017). Identify Key Sequence Features to Improve CRISPR sgRNA Efficacy. *IEEE Access* 5, 26582–26590. doi:10.1109/access.2017.2775703

Chen, W., Chen, L., and Dai, Q. (2021). iMPT-FDNPL: Identification of Membrane Protein Types with Functional Domains and a Natural Language Processing Approach. *Comput. Math. Methods Med.* 2021, 7681497. doi:10.1155/2021/7681497

Choi, J. H., Choi, S.-S., Kim, E. S., Jedrychowski, M. P., Yang, Y. R., Jang, H.-J., et al. (2014). Thrap3 Docks on Phosphoserine 273 of PPARγ and Controls Diabetic Gene Programming. *Genes Dev.* 28, 2361–2369. doi:10.1101/gad.249367.114

Clocquet, A. R., Egan, J. M., Stoffers, D. A., Muller, D. C., Wideman, L., Chin, G. A., et al. (2000). Impaired Insulin Secretion and Increased Insulin Sensitivity in Familial Maturity-Onset Diabetes of the Young 4 (Insulin Promoter Factor 1 Gene). *Diabetes* 49, 1856–1864. doi:10.2337/diabetes.49.11.1856

Cohen, W. W. (1995). "Fast Effective Rule Induction," in Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA, July 9–July 12, 1995, 115–123. doi:10.1016/b978-1-55860-377-6.50023-2

Cornelis, M. C., Rimm, E. B., Curhan, G. C., Kraft, P., Hunter, D. J., Hu, F. B., et al. (2014). Obesity Susceptibility Loci and Uncontrolled Eating, Emotional Eating and Cognitive Restraint Behaviors in Men and Women. *Obesity* 22, E135–E141. doi:10.1002/oby.20592

Cortes, C., and Vapnik, V. (1995). Support-vector Networks. *Mach. Learn* 20, 273–297. doi:10.1007/bf00994018

Cover, T., and Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi:10.1109/tit.1967.1053964

Ding, L., Fan, L., Xu, X., Fu, J., and Xue, Y. (2019). Identification of Core Genes and Pathways in Type 2 Diabetes Mellitus by Bioinformatics Analysis. *Mol. Med. Rep.* 20, 2597–2608. doi:10.3892/mmr.2019.10522

Ding, S., Wang, D., Zhou, X., Chen, L., Feng, K., Xu, X., et al. (2022). Predicting Heart Cell Types by Using Transcriptome Profiles and a Machine Learning Method. *Life* 12, 228. doi:10.3390/life12020228

Disease and Injury Incidence and Prevalence Collaborators (2017). Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 328 Diseases and Injuries for 195 Countries, 1990-2016: a Systematic Analysis for the Global Burden of Disease Study 2016. *Lancet* 390, 1211–1259. doi:10.1016/S0140-6736(17)32154-2

Donath, M. Y., Størling, J., Maedler, K., and Mandrup-Poulsen, T. (2003). Inflammatory Mediators and Islet beta-cell Failure: a Link between Type 1 and Type 2 Diabetes. *J. Mol. Med.* 81, 455–470. doi:10.1007/s00109-003-0450-y

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2008). Monte Carlo Feature Selection for Supervised Classification. *Bioinformatics* 24, 110–117. doi:10.1093/bioinformatics/btm486

Eckardt, K., Görgens, S. W., Raschke, S., and Eckel, J. (2014). Myokines in Insulin Resistance and Type 2 Diabetes. *Diabetologia* 57, 1087–1099. doi:10.1007/s00125-014-3224-x

Elango, S., Venugopal, S., Thangaraj, K., and Viswanadha, V. P. (2014). Novel Mutations in ATPase 8, ND1 and ND5 Genes Associated with Peripheral Neuropathy of Diabetes. *Diabetes Res. Clin. Pract.* 103, e49–e52. doi:10.1016/j.diabres.2013.12.015

Ferrannini, E. (2009). Insulin Resistance versus β-cell Dysfunction in the Pathogenesis of Type 2 Diabetes. *Curr. Diab Rep.* 9, 188–189. doi:10.1007/s11892-009-0031-8

Global Burden of Disease Cancer CollaborationFitzmaurice, C., Allen, C., Barber, R. M., Barregard, L., Bhutta, Z. A., et al. (2017). Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-Years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol.* 3, 524–548. doi:10.1001/jamaoncol.2016.5688

Forst, T., Anastassiadis, E., Diessel, S., Löffler, A., and Pfützner, A. (2014). Effect of Linagliptin Compared with Glimepiride on Postprandial Glucose Metabolism, Islet Cell Function and Vascular Function Parameters in Patients with Type 2 Diabetes Mellitus Receiving Ongoing Metformin Treatment. *Diabetes Metab. Res. Rev.* 30, 582–589. doi:10.1002/dmrr.2525

Gambin, T., Yuan, B., Bi, W., Liu, P., Rosenfeld, J. A., Coban-Akdemir, Z., et al. (2017). Identification of Novel Candidate Disease Genes from De Novo Exonic Copy Number Variants. *Genome Med.* 9, 83. doi:10.1186/s13073-017-0472-7

Gao, H. X., Regier, E. E., and Close, K. L. (2016). International Diabetes Federation World Diabetes Congress 2015. *J. Diabetes* 8, 300. doi:10.1111/1753-0407.12377

Garufi, A., Pistritto, G., Baldari, S., Toietta, G., Cirone, M., and D'Orazi, G. (2017). p53-Dependent PUMA to DRAM Antagonistic Interplay as a Key Molecular Switch in Cell-Fate Decision in Normal/high Glucose Conditions. *J. Exp. Clin. Cancer Res.* 36, 126. doi:10.1186/s13046-017-0596-z

Golubnitschaja, O., Moenkemann, H., Trog, D. B., Blom, H. J., and De Vriese, A. S. (2006). Activation of Genes Inducing Cell-Cycle Arrest and of Increased DNA Repair in the Hearts of Rats with Early Streptozotocin-Induced Diabetes Mellitus. *Med. Sci. Monit.* 12, BR68–74.

Good, A. L., Haemmerle, M. W., Oguh, A. U., Doliba, N. M., and Stoffers, D. A. (2019). Metabolic Stress Activates an ERK/hnRNPK/DDX3X Pathway in Pancreatic β Cells. *Mol. Metab.* 26, 45–56. doi:10.1016/j.molmet.2019.05.009

Guo, X., Zhang, Z., Gerstein, M. B., and Zheng, D. (2009). Small RNAs Originated from Pseudogenes: Cis- or Trans-acting? *PLoS Comput. Biol.* 5, e1000449. doi:10.1371/journal.pcbi.1000449

Houštek, J., Hejzlarová, K., Vrbacký, M., Drahota, Z., Landa, V., Zídek, V., et al. (2012). Nonsynonymous Variants in Mt-Nd2, Mt-Nd4, and Mt-Nd5 Are Linked to Effects on Oxidative Phosphorylation and Insulin Sensitivity in Rat Conplastic Strains. *Physiol. Genomics* 44, 487–494. doi:10.1152/physiolgenomics.00156.2011

Jia, C., Zuo, Y., and Zou, Q. (2018). O-GlcNAcPRED-II: an Integrated Classification Algorithm for Identifying O-GlcNAcylation Sites Based on Fuzzy Undersampling and a K-Means PCA Oversampling Technique. *Bioinformatics* 34, 2029–2036. doi:10.1093/bioinformatics/bty039

Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-Based Machine Learning Model for Predicting the Metabolic Pathways of Compounds. *IEEE Access* 8, 130687–130696. doi:10.1109/access.2020.3009439

Johannes, F., and Widmer, G. (1994). "Incremental Reduced Error Pruning," in Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, July 10–July 13, 1994, 70–77. doi:10.1016/b978-1-55860-335-6.50017-9

Johnson, D. S. (1974). Approximation Algorithms for Combinatorial Problems. *J. Comput. Syst. Sci.* 9, 256–278. doi:10.1016/s0022-0000(74)80044-9

Kahn, S. E. (2003). The Relative Contributions of Insulin Resistance and Beta-Cell Dysfunction to the Pathophysiology of Type 2 Diabetes. *Diabetologia* 46, 3–19. doi:10.1007/s00125-002-1009-0

Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D. R., Wu, Y.-M., Cao, X., et al. (2012). Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers. *Cell* 149, 1622–1634. doi:10.1016/j.cell.2012.04.041

Kandaswamy, K. K., Chou, K.-C., Martinetz, T., Möller, S., Suganthan, P. N., Sridharan, S., et al. (2011). AFP-pred: A Random Forest Approach for Predicting Antifreeze Proteins from Sequence-Derived Properties. *J. Theor. Biol.* 270, 56–62. doi:10.1016/j.jtbi.2010.10.037

Kim, J.-H., Choi, S.-Y., Kang, B.-H., Lee, S.-M., Park, H. S., Kang, G.-Y., et al. (2013). AMP-activated Protein Kinase Phosphorylates CtBP1 and Down-Regulates its Activity. *Biochem. Biophysical Res. Commun.* 431, 8–13. doi:10.1016/j.bbrc.2012.12.117

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in International Joint Conference on Artificial Intelligence, Montreal Quebec Canada, August 20–August 25, 1995 (Lawrence Erlbaum Associates Ltd), 1137–1145.

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., et al. (2017). Single-cell Transcriptomes Identify Human Islet Cell Signatures and Reveal Cell-type-specific Expression Changes in Type 2 Diabetes. *Genome Res.* 27, 208–222. doi:10.1101/gr.212720.116

Lebovitz, C. B., Robertson, A. G., Goya, R., Jones, S. J., Morin, R. D., Marra, M. A., et al. (2015). Cross-cancer Profiling of Molecular Alterations within the Human Autophagy Interaction Network. *Autophagy* 11, 1668–1687. doi:10.1080/15548627.2015.1067362

Lee, M.-S. (2014). Role of Islet β Cell Autophagy in the Pathogenesis of Diabetes. *Trends Endocrinol. Metabolism* 25, 620–627. doi:10.1016/j.tem.2014.08.005

Li, X., Lu, L., Lu, L., and Chen, L. (2022). Identification of Protein Functions in Mouse with a Label Space Partition Method. *Mbe* 19, 3820–3842. doi:10.3934/mbe.2022176

Li, Z., Wang, D., Liao, H., Zhang, S., Guo, W., Chen, L., et al. (2022). Exploring the Genomic Patterns in Human and Mouse Cerebellums via Single-Cell Sequencing and Machine Learning Method. *Front. Genet.* 13, 857851. doi:10.3389/fgene.2022.857851

Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of Drug Side Effects with a Refined Negative Sample Selection Strategy. *Comput. Math. Methods Med.* 2020, 1573543. doi:10.1155/2020/1573543

Liu, H., Hu, B., Chen, L., and Lu, L. (2021). Identifying Protein Subcellular Location with Embedding Features Learned from Networks. *Cp* 18, 646–660. doi:10.2174/1570164617999201124142950

Liu, H., and Setiono, R. (1998). Incremental Feature Selection. *Appl. Intell.* 9, 217–230. doi:10.1023/a:1008363719778

Liu, J.-Y., Yao, J., Li, X.-M., Song, Y.-C., Wang, X.-Q., Li, Y.-J., et al. (2014). Pathogenic Role of lncRNA-MALAT1 in Endothelial Cell Dysfunction in Diabetes Mellitus. *Cell Death Dis.* 5, e1506. doi:10.1038/cddis.2014.466

Lontchi-Yimagou, E., Sobngwi, E., Matsha, T. E., and Kengne, A. P. (2013). Diabetes Mellitus and Inflammation. *Curr. Diab Rep.* 13, 435–444. doi:10.1007/s11892-013-0375-y

Lyons, T. J., and Basu, A. (2012). Biomarkers in Diabetes: Hemoglobin A1c, Vascular and Tissue Markers. *Transl. Res.* 159, 303–312. doi:10.1016/j.trsl.2012.01.009

Ma, L., and Zheng, J. (2018). Single-cell Gene Expression Analysis Reveals β-cell Dysfunction and Deficit Mechanisms in Type 2 Diabetes. *BMC Bioinforma.* 19, 515. doi:10.1186/s12859-018-2519-1

Mannino, G. C., and Sesti, G. (2012). Individualized Therapy for Type 2 Diabetes: Clinical Implications Of Pharmacogenetic Data. *Mol. Diagn Ther.* 16, 285–302. doi:10.1007/s40291-012-0002-7

Marques, Y. B., De Paiva Oliveira, A., Ribeiro Vasconcelos, A. T., and Cerqueira, F. R. (2016). Mirnacle: Machine Learning with SMOTE and Random Forest for Improving Selectivity in Pre-miRNA Ab Initio Prediction. *BMC Bioinforma.* 17, 474. doi:10.1186/s12859-016-1343-8

Mathews, C. E., Leiter, E. H., Spirina, O., Bykhovskaya, Y., Gusdon, A. M., Ringquist, S., et al. (2005). mt-Nd2 Allele of the ALR/Lt Mouse Confers Resistance against Both Chemically Induced and Autoimmune Diabetes. *Diabetologia* 48, 261–267. doi:10.1007/s00125-004-1644-8

Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica Biophysica Acta (BBA) - Protein Struct.* 405, 442–451. doi:10.1016/0005-2795(75)90109-9

Mella, M. T., Kohari, K., Jones, R., Peña, J., Ferrara, L., Stone, J., et al. (2016). Mitochondrial Gene Expression Profiles Are Associated with Intrahepatic Cholestasis of Pregnancy. *Placenta* 45, 16–23. doi:10.1016/j.placenta.2016.07.002

Mirza, A. H., Berthelsen, C. H., Seemann, S. E., Pan, X., Frederiksen, K. S., Vilien, M., et al. (2015). Transcriptomic Landscape of lncRNAs in Inflammatory Bowel Disease. *Genome Med.* 7, 39. doi:10.1186/s13073-015-0162-2

Onesime, M., Yang, Z., and Dai, Q. (2021). Genomic Island Prediction via Chi-Square Test and Random Forest Algorithm. *Comput. Math. Methods Med.* 2021, 9969751. doi:10.1155/2021/9969751

Pan, X.-Y., and Shen, H.-B. (2009). Robust Prediction of B-Factor Profile from Sequence Using Two-Stage SVR Based on Random Forest Feature Selection. *Ppl* 16, 1447–1454. doi:10.2174/092986609789839250

Pan, X., Li, H., Zeng, T., Li, Z., Chen, L., Huang, T., et al. (2021). Identification of Protein Subcellular Localization with Network and Functional Embeddings. *Front. Genet.* 11, 626500. doi:10.3389/fgene.2020.626500

Pandey, A., Chawla, S., and Guchhait, P. (2015). Type-2 Diabetes: Current Understanding and Future Perspectives. *IUBMB Life* 67, 506–513. doi:10.1002/iub.1396

Poliseno, L., Marranci, A., and Pandolfi, P. P. (2015). Pseudogenes in Human Cancer. *Front. Med.* 2, 68. doi:10.3389/fmed.2015.00068

Prentki, M., and Nolan, C. J. (2006). Islet Cell Failure in Type 2 Diabetes. *J. Clin. Investigation* 116, 1802–1812. doi:10.1172/jci29103

Pullen, T. J., and Rutter, G. A. (2014). Roles of lncRNAs in Pancreatic Beta Cell Identity and Diabetes Susceptibility. *Front. Genet.* 5, 193. doi:10.3389/fgene.2014.00193

Quinlan, J. R. (1990). Learning Logical Definitions from Relations. *Mach. Learn* 5, 239–266. doi:10.1007/bf00117105

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. *Cell Metab.* 24, 593–607. doi:10.1016/j.cmet.2016.08.020

Siddiqui, A., Madhu, S. V., Sharma, S. B., and Desai, N. G. (2015). Endocrine Stress Responses and Risk of Type 2 Diabetes Mellitus. *Stress* 18, 498–506. doi:10.3109/10253890.2015.1067677

Soto, A. G., Cheruvu, S., Bialo, D., and Quintos, J. B. (2014). Refractory Diabetes Insipidus Leading to Diagnosis of Type 2 Diabetes Mellitus and Non-ketotic Hyperglycemia in an Adolescent Male. *R. I. Med. J. (2013)* 97, 34–35.

Stancakova, A., and Laakso, M. (2016). Genetics of Type 2 Diabetes. *Endocr. Dev.* 31, 203–220. doi:10.2337/dc10-1013

Tang, S., and Chen, L. (2022). iATC-NFMLP: Identifying Classes of Anatomical Therapeutic Chemicals Based on Drug Networks, Fingerprints and Multilayer Perceptron. *Curr. Bioinforma.* 17. doi:10.2174/1574893617666220318093000

Tao, Z., Shi, A., and Zhao, J. (2015). Epidemiological Perspectives of Diabetes. *Cell Biochem. Biophys.* 73, 181–185. doi:10.1007/s12013-015-0598-4

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The Multilayered Complexity of ceRNA Crosstalk and Competition. *Nature* 505, 344–352. doi:10.1038/nature12986

Taylor, R. (2013). Type 2 Diabetes: Etiology And Reversibility. *Diabetes Care* 36, 1047–1055. doi:10.2337/dc12-1805

Thorens, B. (2014). Neural Regulation of Pancreatic Islet Cell Mass and Function. *Diabetes Obes. Metab.* 16 (Suppl. 1), 87–95. doi:10.1111/dom.12346

Torrell, H., Montaña, E., Abasolo, N., Roig, B., Gaviria, A. M., Vilella, E., et al. (2013). Mitochondrial DNA (mtDNA) in Brain Samples from Patients with Major Psychiatric Disorders: Gene Expression Profiles, mtDNA Content and Presence of the mtDNA Common Deletion. *Am. J. Med. Genet.* 162, 213–223. doi:10.1002/ajmg.b.32134

Trujillo, J. M., and Nuffer, W. (2014). GLP-1 Receptor Agonists for Type 2 Diabetes Mellitus: Recent Developments and Emerging Agents. *Pharmacotherapy* 34, 1174–1186. doi:10.1002/phar.1507

Tsai, A., Cowan, M. R., Johnson, D. G., and Brannon, P. M. (1994). Regulation of Pancreatic Amylase and Lipase Gene Expression by Diet and Insulin in Diabetic Rats. *Am. J. Physiology-Gastrointestinal Liver Physiology* 267, G575–G583. doi:10.1152/ajpgi.1994.267.4.g575

Tseng, C. H., Chen, C. J., and Landolph, J. R., Jr. (2012). Diabetes and Cancer: Epidemiological, Clinical, and Experimental Perspectives. *Exp. Diabetes Res.* 2012, 101802. doi:10.1155/2012/101802

Tsuyuki, S., Takabayashi, M., Kawazu, M., Kudo, K., Watanabe, A., Nagata, Y., et al. (2014). Detection ofWIPI1mRNA as an Indicator of Autophagosome Formation. *Autophagy* 10, 497–513. doi:10.4161/auto.27419

Urbanová, M., Mráz, M., Ďurovcová, V., Trachta, P., Kloučková, J., Kaválková, P., et al. (2017). The Effect of Very-Low-Calorie Diet on Mitochondrial Dysfunction in Subcutaneous Adipose Tissue and Peripheral Monocytes of Obese Subjects with Type 2 Diabetes Mellitus. *Physiol. Res.* 66, 811–822. doi:10.33549/physiolres.933469

Walker, W. E., Kurscheid, S., Joshi, S., Lopez, C. A., Goh, G., Choi, M., et al. (2015). Increased Levels of Macrophage Inflammatory Proteins Result in Resistance to R5-Tropic HIV-1 in a Subset of Elite Controllers. *J. Virol.* 89, 5502–5514. doi:10.1128/jvi.00118-15

Wang, Y., Xu, Y., Yang, Z., Liu, X., and Dai, Q. (2021). Using Recursive Feature Selection with Random Forest to Improve Protein Structural Class Prediction for Low-Similarity Sequences. *Comput. Math. Methods Med.* 2021, 5529389. doi:10.1155/2021/5529389

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2018). Exploring Sequence-Based Features for the Improved Prediction of DNA N4-Methylcytosine Sites in Multiple Species. *Bioinformatics* 35, 1326–1333. doi:10.1093/bioinformatics/bty824

Westermark, G. T., and Westermark, P. (2008). Importance of Aggregated Islet Amyloid Polypeptide for the Progressive Beta-Cell Failure in Type 2 Diabetes and in Transplanted Human Islets. *Exp. Diabetes Res.* 2008, 528354. doi:10.1155/2008/528354

Wu, Z., and Chen, L. (2022). Similarity-based Method with Multiple-Feature Sampling for Predicting Drug Side Effects. *Comput. Math. Methods Med.* 2022, 1–13. doi:10.1155/2022/9547317

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., et al. (2016). RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metab.* 24, 608–615. doi:10.1016/j.cmet.2016.08.018

Yabe, D., Seino, Y., Fukushima, M., and Seino, S. (2015). β Cell Dysfunction versus Insulin Resistance in the Pathogenesis of Type 2 Diabetes in East Asians. *Curr. Diab Rep.* 15, 602. doi:10.1007/s11892-015-0602-9

Yang, Y., and Chen, L. (2022). Identification of Drug-Disease Associations by Using Multiple Drug and Disease Networks. *Cbio* 17, 48–59. doi:10.2174/1574893616666210825115406

Zhang, Y.-H., Li, H., Zeng, T., Chen, L., Li, Z., Huang, T., et al. (2021a). Identifying Transcriptomic Signatures and Rules for SARS-CoV-2 Infection. *Front. Cell Dev. Biol.* 8, 627302. doi:10.3389/fcell.2020.627302

Zhang, Y.-H., Li, Z., Zeng, T., Chen, L., Li, H., Huang, T., et al. (2021b). Detecting the Multiomics Signatures of Factor-specific Inflammatory Effects on Airway Smooth Muscles. *Front. Genet.* 11, 599970. doi:10.3389/fgene.2020.599970

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021c). Determining Protein-Protein Functional Associations by Functional Rules Based on Gene Ontology and KEGG Pathway. *Biochimica Biophysica Acta (BBA) - Proteins Proteomics* 1869, 140621. doi:10.1016/j.bbapap.2021.140621

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting Drug Side Effects with Compact Integration of Heterogeneous Networks. *Cbio* 14, 709–720. doi:10.2174/1574893614666190220114644

Zhao, X., Chen, L., and Lu, J. (2018). A Similarity-Based Method for Prediction of Drug Side Effects with Heterogeneous Information. *Math. Biosci.* 306, 136–144. doi:10.1016/j.mbs.2018.09.010

Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020b). iATC-FRAKEL: a Simple Multi-Label Web Server for Recognizing Anatomical Therapeutic Chemical Classes of Drugs with Their Fingerprints Only. *Bioinformatics* 36, 3568–3569. doi:10.1093/bioinformatics/btaa166

Zhou, J. P., Chen, L., and Guo, Z. H. (2020a). iATC-NRAKEL: An Efficient Multi-Label Classifier for Recognizing Anatomical Therapeutic Chemical Classes of Drugs. *Bioinformatics* 36, 1391–1396. doi:10.1093/bioinformatics/btz757

Zhou, X., Hao, Q., Liao, J., Zhang, Q., and Lu, H. (2013). Ribosomal Protein S14 Unties the MDM2-P53 Loop upon Ribosomal Stress. *Oncogene* 32, 388–396. doi:10.1038/onc.2012.63

Zhu, Y., Hu, B., Chen, L., and Dai, Q. (2021). iMPTCE-Hnetwork: A Multilabel Classifier for Identifying Metabolic Pathway Types of Chemicals and Enzymes with a Heterogeneous Network. *Comput. Math. Methods Med.* 2021, 6683051. doi:10.1155/2021/6683051

Zick, Y. (2001). Insulin Resistance: a Phosphorylation-Based Uncoupling of Insulin Signaling. *Trends Cell Biol.* 11, 437–441. doi:10.1016/s0962-8924(01)81297-6

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 | Frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 13 July 2022
doi: 10.3389/fbioe.2022.917086

Deconvolution of a Large Cohort of Placental Microarray Data Reveals Clinically Distinct Subtypes of Preeclampsia

Tian Yao[1,2†], Qiming Liu[1†] and Weidong Tian[1,3,4]*

[1]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Department of Computational Biology, School of Life Sciences, Fudan University, Shanghai, China, [2]Human Phenome Institute, Fudan University, Shanghai, China, [3]Children's Hospital of Fudan University, Shanghai, China, [4]Qilu Children's Hospital of Shandong University, Jinan, China

It has been well established that the dysfunctional placenta plays an important role in the pathogenesis of preeclampsia (PE), a hypertensive disorder in pregnancy. However, it is not well understood how individual cell types in the placenta are involved in placenta dysfunction because of limited single-cell studies of placenta with PE. Given that a high-resolution single-cell atlas in the placenta is now available, deconvolution of publicly available bulk PE transcriptome data may provide us with the opportunity to investigate the contribution of individual placental cell types to PE. Recent benchmark studies on deconvolution have provided suggestions on the strategy of marker gene selection and the choice of methodologies. In this study, we experimented with these suggestions by using real bulk data with known cell-type proportions and established a deconvolution pipeline using CIBERSORT. Applying the deconvolution pipeline to a large cohort of PE placental microarray data, we found that the proportions of trophoblast cells in the placenta were significantly different between PE and normal controls. We then predicted cell-type-level expression profiles for each sample using CIBERSORTx and found that the activities of several canonical PE-related pathways were significantly altered in specific subtypes of trophoblasts in PE. Finally, we constructed an integrated expression profile for each PE sample by combining the predicted cell-type-level expression profiles of several clinically relevant placental cell types and identified four clusters likely representing four PE subtypes with clinically distinct features. As such, our study showed that deconvolution of a large cohort of placental microarray provided new insights about the molecular mechanism of PE that would not be obtained by analyzing bulk expression profiles.

Keywords: deconvolution, preeclampsia, heterogeneity, single-cell, pipeline

OPEN ACCESS

Edited by:
Zhaoyuan Fang,
Zhejiang University, China

Reviewed by:
Xiaoxiao Sun,
University of Arizona, United States
Lingling An,
University of Arizona, United States
Ning Shen,
Zhejiang University, China

****Correspondence:***
Weidong Tian
weidong.tian@fudan.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *10 April 2022*
Accepted: *13 June 2022*
Published: *13 July 2022*

Citation:
Yao T, Liu Q and Tian W (2022) Deconvolution of a Large Cohort of Placental Microarray Data Reveals Clinically Distinct Subtypes of Preeclampsia. Front. Bioeng. Biotechnol. 10:917086. doi: 10.3389/fbioe.2022.917086

1 INTRODUCTION

Preeclampsia (PE) is a hypertensive disorder of pregnancy and is the main reason for maternal and fetal morbidity and mortality (Bokslag et al., 2016). Abnormal development and dysfunction of the placenta are thought to be the main cause of PE though detailed pathophysiology is still not fully understood (Horii et al., 2021). As the placenta is a heterogeneous tissue consisting of diverse types of cells, single-cell studies of PE's placentas are expected to lead to a better understanding of the molecular mechanisms underlining PE pathogenesis. However, most PE transcriptome studies published so far were done at the bulk level (Leavey et al., 2016; Robineau-Charette et al., 2020; Yadama et al., 2020; Xu et al., 2021). A recently published single-cell study on the placenta of PE included only three samples each in the PE and the control groups (Zhang et al., 2021), providing a limited number of samples to investigate the association of individual cell types in the placenta with PE. Cell-type deconvolution is a technology that can infer cell-type proportions from bulk transcription profiles when cell-type-specific expression profiles of marker genes are available (Jin and Liu, 2021). Given that the high-resolution single-cell atlas of the placenta is now available (Suryawanshi et al., 2018; Vento-Tormo et al., 2018), reanalyzing existing bulk PE transcriptome data by deconvolution may therefore provide us with the opportunity to investigate the contribution of individual placental cell types in the placenta to PE.

Numerous deconvolution methods have been developed (Newman et al., 2015; Hao et al., 2019; Newman et al., 2019; Tsoucas et al., 2019; Wang et al., 2019; Dong et al., 2021), and they can be generally divided into two broad categories (Cobos et al., 2020): the bulk and the single-cell reference-based methods, respectively, with the former requiring a predefined cell-type-specific signature gene matrix and the latter not. CIBERSORT (Newman et al., 2015) and CIBERSORTx (Newman et al., 2019) are the representative methods of these two categories, respectively. The use of deconvolution methods has greatly accelerated the study of diseases. For example, prognostic biomarkers of renal cell carcinoma were identified by estimating the proportions of tumor-infiltrating immune cells by cell-type deconvolution using CIBERSORT (Zhang et al., 2019). Recent benchmark studies evaluating the performance of current deconvolution methods (Cobos et al., 2020; Jin and Liu, 2021; Nadel et al., 2021) have provided suggestions on the strategy of marker gene selection and the choice of deconvolution methodologies. For our study, i.e., conducting deconvolution on bulk PE transcriptome data, however, on the one hand, the detailed thresholds for marker selection need to be specified. On the other hand, we still need to decide on one of several recommended methods to perform deconvolution.

In this study, we followed the strategy suggested by Francisco et al. (Cobos et al., 2020) to determine the thresholds for marker gene selection. Then, by using different sources (RNA-seq and microarray) of real bulk data with known cell-type proportions, we evaluated several deconvolution methods recommended by Francisco et al. using two measures—the Pearson correlation coefficient between the predicted and true cell-type proportions (PCC_P) and the Pearson correlation coefficient between the predicted and true bulk transcripts (PCC_T). As PCC_T can be directly calculated from a deconvolution, it has been suggested to be potentially useful for improving the performance of deconvolution (Newman et al., 2015; Dong et al., 2021). We, therefore, investigated the relationships between the two PCCs to explore the possibility of using PCC_T to select a deconvolution method. Finally, we applied the deconvolution pipeline derived from the above-described experiments to a large cohort of PE microarray data that have detailed clinical phenotypes (Leavey et al., 2016). We then conducted an in-depth analysis on the deconvolution results and particularly explored the cell-type-level expression profiles predicted based on the estimated placental cell-type proportions. Our results led to four PE subtypes with clinically distinct features that would not be observed by analyzing bulk gene expression profiles.

2 RESULTS

2.1 The Development of a Practical Pipeline for the Deconvolution of Placenta Microarray Data

The benchmark study by Francisco et al. (Cobos et al., 2020) provided suggestions on marker gene selection and the choices of methodologies. For marker gene selection, it is recommended to use a stringent selection strategy by using the following three measures—logFC, logCPM, and SecondFC, representing the cell-type-specificity across all cell types, the averaged expression level across all cell types, and the cell-type to cell-type difference of a marker gene, respectively (see **Section 3** for details about the definition of these three measures). For the choice of methodologies, it recommended several bulk reference-based methods, including CIBERSORT (Newman et al., 2015), robust linear regression (RLR) (Venables and Ripley, 2002), FARDEEP (Hao et al., 2019), OLS (Chambers et al., 1990), and nonnegative least squares (NNLS) (Katharine et al., 2012), and several single-cell reference-based methods, including DWLS (Tsoucas et al., 2019), MuSiC (Wang et al., 2019), and SCDC (Dong et al., 2021). We added CIBERSORTx (Newman et al., 2019), which is based on CIBERSORT's improved method of using single-cell data as input. There are also nonreference-based deconvolution methods available, such as ssFrobenius (Gaujoux and Seoighe, 2012). However, Avila Cobos et al. (2018) had shown that reference-based methods would work better than nonreference-based methods when the reference expression profiles are available. Because the single-cell reference of the placental atlas is available in this study, we did not consider the nonreference-based deconvolution methods in this study. Although the above suggestions were useful, in our case, we still need to determine the thresholds for the three marker gene selection measures and also have to choose a method from the recommended ones.

To determine the thresholds for marker gene selection, we selected the peripheral blood mononuclear cells (PBMCs) bulk data produced by Finotello et al. (2019) in which cell-type proportions were determined by flow cytometry for deconvolution. We then obtained the reference expression profiles of the immune cell types from the RNA-seq data generated by Hoek et al. (2015) to generate the signature gene matrix. We fixed the thresholds of both logFC and log CPM to be one and experimented with different thresholds of SecondFC to construct the signature gene matrices. We used the Pearson correlation coefficient between the predicted and true cell-type proportions (PCC_P) for evaluating the performance of deconvolution. We found that with the increase of SecondFC, the average correlation between cell types in the signature gene matrix decreases, but PCC_P increases; when the similarity

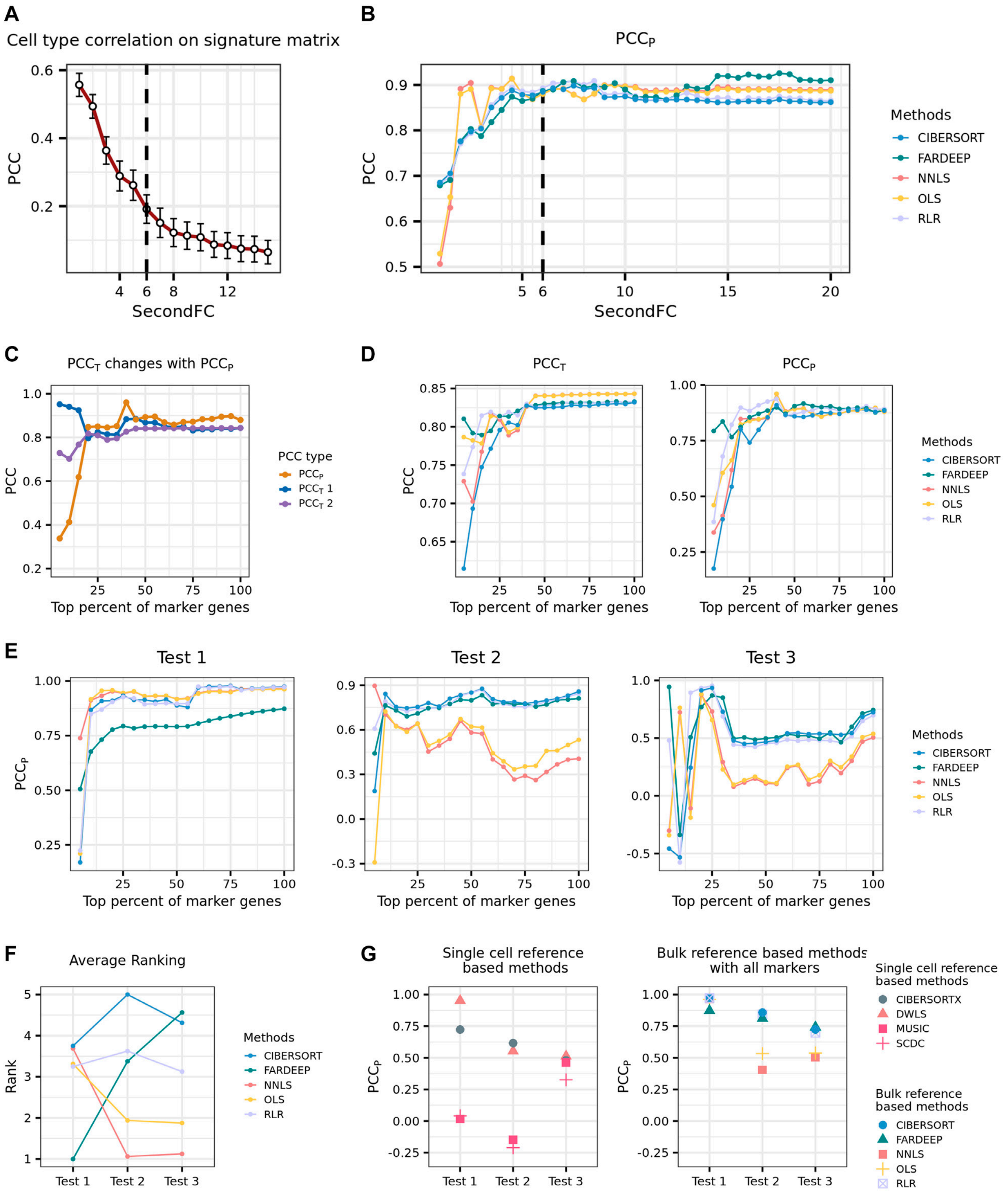


FIGURE 1 | Development of a practical pipeline for the deconvolution of placenta microarray data. **(A)** Average PCC on PBMC signature matrix changing with SecondFC cutoff. **(B)** PCC_p of different methods changing with SecondFC cutoff. **(C)** The changes of PCC_T and PCC_p, where the predicted expression profiles of the former and the latter were computed by using the input signature gene matrix varied. PCC_{T1} and PCC_{T2} are the PCC between the predicted and the real bulk expression profiles on inputted signature gene matrix and the signature gene matrix with all marker genes, respectively. **(D)** The changes of PCC_T and PCC_p by using different deconvolution methods, where PCC_T refers to PCC_{T2} in **(C)**. **(E)** Three benchmark tests to evaluate the performance of different deconvolution methods. In Tests 1 and 2, the reference expression profiles were from the 10X scRNA-seq PBMC data generated by Ding et al. (2020), and the bulk data were Finotello's PBMC RNA-seq data and Newman's PBMC microarray data. In Test 3, the bulk data were the same as in Test 2, while the reference expression profiles were the Drop-seq and inDrops scRNA-seq PBMC data generated by Ding et al. (2020) **(F)** The average rank of different deconvolution methods across the three tests in **(E)**. **(G)** The comparison of the performance of single-cell and bulk reference-based methods across the three tests in **(E)**.

decreases to an inflection point, PCC_P would reach a high level (**Figures 1A,B**). Accordingly, the threshold of SecondFC could be determined by investigating the relationship between SecondFC and the average correlation between cell types in the signature gene matrix.

Next, we aimed to determine which deconvolution method should be used in practice. Given the estimated cell-type proportions by a deconvolution method, the predicted expression profiles of bulk transcripts can be computed by $T = C \cdot P$, where T represents the predicted bulk expression profile, C is the signature gene matrix, and P is the estimated cell-type proportions. The PCC between the predicted and true expression of bulk transcripts (PCC_T) can then be calculated. It is assumed that the closer the estimated cell-type proportions to true cell-type proportions, that is, a higher PCC_P, the closer the predicted expression of bulk transcripts to true expression, that is, a higher PCC_T. It has thus been proposed that maximizing PCC_T may have the effect of maximizing PCC_P (Newman et al., 2015; Dong et al., 2021). If this were true, then PCC_T may also be used for selecting the deconvolution method, that is, a method with a greater PCC_T ought to have a greater PCC_P. To test this possibility, here we investigated the relationships between PCC_p and PCC_T.

From the marker genes selected by following the above-described parameters, we selected a top fraction of genes according to their logFC to generate a signature gene matrix and conduct deconvolution. A pair of PCC_p and PCC_T could be calculated for each selected fraction of marker genes, and a series of paired PCC_p and PCC_T could be calculated by increasing the fraction of marker genes. Note that there are two ways of predicting T: one in which C is the signature gene matrix corresponding to a selected fraction of marker genes and varies when the fraction changes, and another in which C is the signature gene matrix corresponding to the whole set of marker genes and does not change with different selected fractions. The PCC_T corresponding to these two situations was named PCC_{T1} and PCC_{T2}, respectively. In general, PCC_p increased with the inclusion of more marker genes, and the increase was relatively sharp before the inclusion of the top 25% of marker genes. Interestingly, before the inclusion of the top 25% of marker genes, PCC_{T1} and PCC_P were negatively correlated, whereas PCC_{T2} and PCC_P were positively correlated (**Figure 1C**). Although for a given method, a higher PCC_{T2} usually indicates a higher PCC_P, this prediction cannot be generalized when the comparison is across different methods (**Figure 1D**). Accordingly, we concluded that it is not possible to select a deconvolution method by comparing their PCC_T.

In our situation of deconvolution, the reference expression profiles were obtained from a single-cell study of the placenta (Vento-Tormo et al., 2018) while the bulk data were from a large cohort of microarray study on PE (Leavey et al., 2016). In order to select a deconvolution method from the recommended ones, we, therefore, prepared three benchmark tests whose degree of deconvolution difficulty was considered to be similar to ours and reasoned that a method performing stably across these three datasets would also likely perform well in our situation. In the first benchmark dataset (Test 1), the bulk data were PBMC RNA-seq data produced by Finotello et al. (2019), and the reference expression profiles were from the single-cell PBMC RNA-seq data generated by Ding et al. (2020) using 10X sequencing platform. In the second benchmark dataset (Test 2), the reference expression profiles were the same as in Test 1, while the bulk data were PBMC microarray data (Newman et al., 2015). In the third benchmark dataset (Test 3), the bulk data were the same as in Test 2, while the reference expression profiles were from the single-cell PBMC RNA-seq data generated by Ding et al. (2020) using Drop-seq and inDrops sequencing platform. In each of the three benchmarks, the signature gene matrices were produced from a top fraction of marker genes selected according to the previously described procedures. In general, most bulk reference-based methods perform better when more marker genes are used, and CIBERSORT and RLR achieved better performance than the other three methods did across the three tests (**Figure 1E**). To further quantify how stable a method's performance is with the inclusion of more marker genes, we ranked the performance of the five methods at a given fraction (from top 25% to top 100%) of marker genes and then calculated the averaged rank of each method. We found that CIBERSORT had the most stable overall performance across the three tests (**Figure 1F**). We also evaluated the performance of four single-cell reference-based methods (DWLS, MuSiC, SCDC, and CIBERSORTx) in these three tests and found that DWLS performed the best among the four methods though its overall performance was worse than CIBERSORT's (**Figure 1G**).

Based on the above analyses, we, therefore, developed a practical pipeline for the deconvolution of PE microarray data. We would follow the procedures described previously to select marker genes and construct a signature gene matrix. Then, we would use CIBERSORT, the method with the most stable and good performance across the three benchmark tests, to perform deconvolution.

2.2 Deconvolution of Preeclampsia Placenta Microarray Data Revealed Significantly Altered Proportions of Trophoblasts in Preeclampsia

The cohort of PE placental microarray data was constructed by Leavey et al. (2016) and included a total number of 330 samples (157 PE and 173 control), of which 157 had detailed clinical information. The clinical information is mainly about the fetal and maternal state, like newborn weight z-score, maximum systolic bp, mode proteinuria, etc. The reference expression profiles were obtained from the single-cell placental RNA-seq data produced by Vento-Tormo et al. (2018). Following Francisco's suggestion to include all cell types that possibly exist in the bulk mixture, we selected the expression profiles of all major cell types (subpopulations were pooled) in the placenta and the blood of the Vento-Tormo dataset (see **Section 3** for details) and constructed a signature matrix consisting of endothelial cells (Endo), epithelial cells (Epi), fibroblasts (FB), three types of trophoblasts cells—villous cytotrophoblasts (VCT), syncytiotrophoblasts (SCT), and extravillous trophoblasts (EVT), and eight types of immune cells—Hofbauer (HB), natural killer

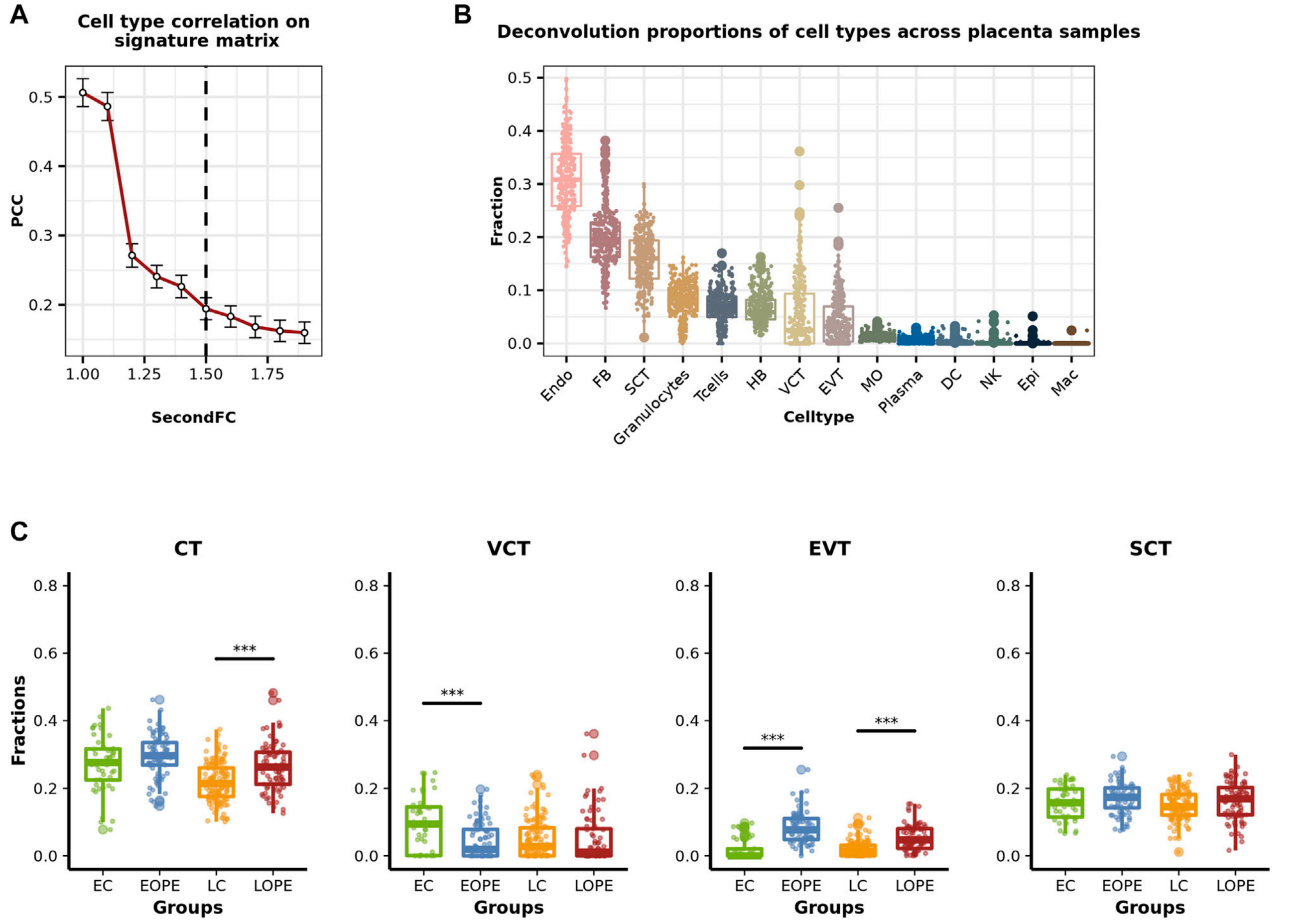


FIGURE 2 | Estimated cell-type proportions of placental samples included in the cohort of placental microarray data. **(A)** Average PCC on placenta signature matrix changing with SecondFC cutoff. **(B)** Boxplots of the estimated proportions of different placental cell types in the cohort of placenta microarray data. **(C)** Comparison of the estimated proportions of trophoblasts between PE and normal samples. CT: cytotrophoblast, EC: early-stage control, EOPE: early-onset preeclampsia, LC: late-stage control, LOPE: late-onset preeclampsia.

(NK), T cells, plasma, granulocytes, monocyte (MO), macrophage (Mac), and dendritic cells (DC). Here, we set SecondFC to 1.5 (**Figure 2A**) by following the above-described procedures to select the marker genes for deconvolution and applied CIBERSORT to perform the deconvolution. The deconvolution results showed that Endo, the major component cells of placental blood vessels, were the largest population of cells in the placenta samples of this cohort, while FB, which is located within the villus core matrix with HB, and SCT were the second and the third largest population of cells, respectively (**Figure 2B**). However, if VCT and EVT were considered together with SCT, then trophoblasts were the largest population of cells in the placenta. Among the eight types of immune cells, however, only granulocytes and T cells accounted for a noticeable proportion in the placental samples (**Figure 2B**).

PE can be generally classified as early-onset PE (EOPE) and late-onset PE (LOPE) depending on the gestational age (GA) (34 weeks) of disease onset (Von Dadelszen et al., 2003). Following this definition, we then classified the PE samples in this cohort as EOPE or LOPE and also classified the normal samples as early control (EC) or late control (LC), respectively. As trophoblasts are the major population of cells in the placenta and are also responsible for the normal function of the placenta, we compared the proportion of trophoblasts between PE and normal controls and observed significant differences (**Figure 2C**). LOPE has a significantly higher proportion of trophoblasts than its group of normal controls (**Figure 2C**). As for the subpopulations of trophoblasts, compared to normal controls, VCT's proportion was significantly lower in EOPE and lower but not significant in LOPE; EVT's proportion was significantly higher in both EOPE and LOPE; SCT's proportion was not significantly altered in PE (**Figure 2C**). It has been shown that the impaired invasive ability of EVT is a major reason for dysfunctional placenta in PE (Crosley et al., 2013). Here, the significantly increased proportion of EVT in PE may be because of a compensatory enhancement of EVT production occurring in response to dysfunctional EVT.

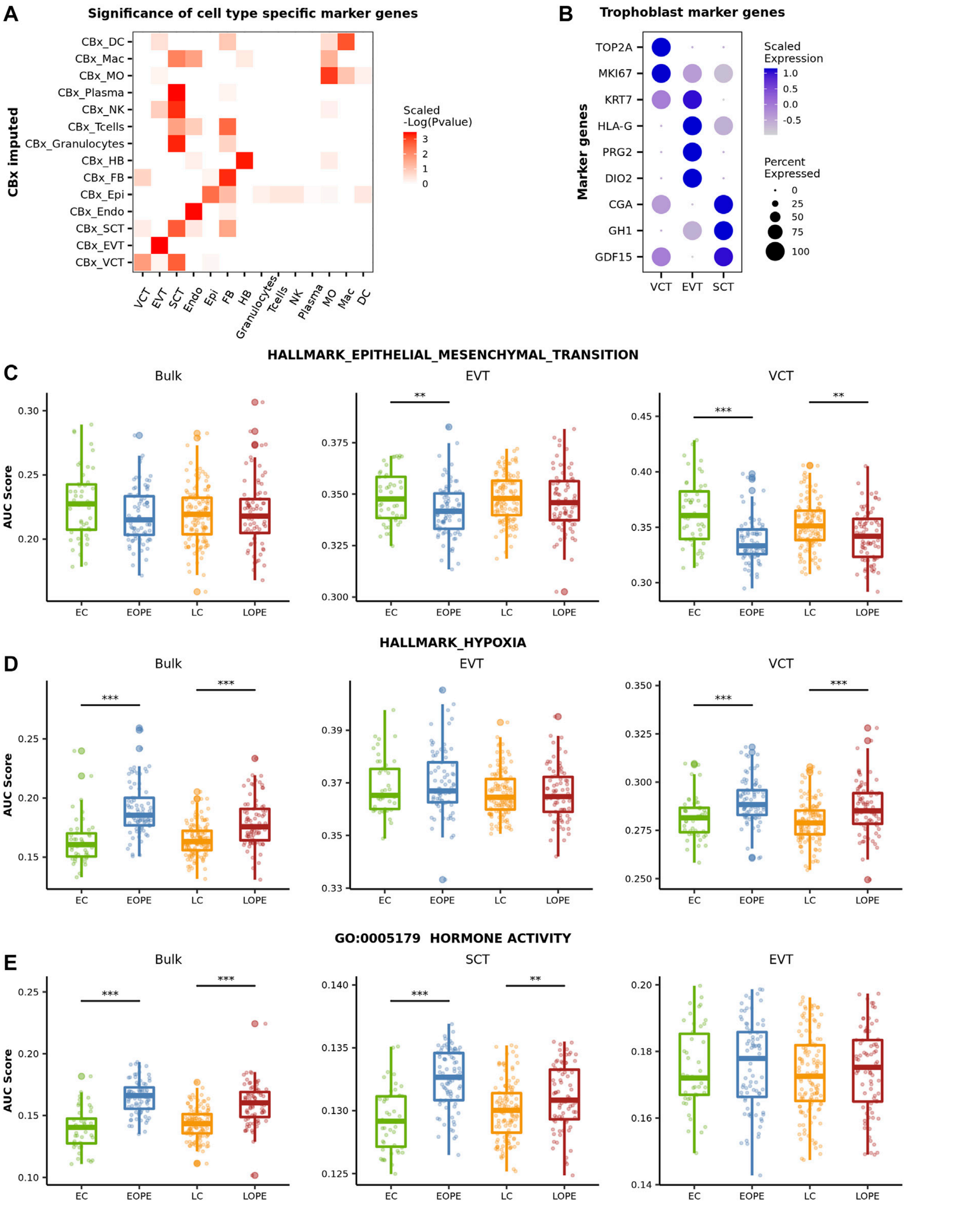


FIGURE 3 | Comparison of the activity of canonical PE-related pathways in between PE and normal samples. **(A)** Assessment of the biological relevance of the predicted cell-type-level expression profiles. We first averaged the expression profile of imputed transcriptome in each cell type. Then, the Wilcox test was used to evaluate if the expression on the averaged profile of the cell-type marker genes is specifically high in the corresponding cell type. The negative log *P* value of the Wilcox test was scaled by rows. **(B)** The expression level of classic trophoblasts marker genes in the predicted cell-type-level expression profiles of three trophoblast subtypes. **(C–E)** Activity of the canonical PE-related pathway in different groups of samples. The activity was measured by AUCell.

2.3 The Predicted Cell-Type-Level Expression Profiles Revealed Patterns of Cell-Type-Specific Gene Expression Alterations in Preeclampsia

Given the estimated cell-type proportions, CIBERSORTx provides a way to infer cell-type-level expression profiles (Steen et al., 2020). Here, we applied the high-resolution mode of CIBERSORTx with the default parameters to predict the expression profiles of placental cell types for each sample. To validate that the predicted cell-type-level expression profiles are biologically meaningful, we tested whether the corresponding cell-type-specific marker genes identified from the reference expression profiles were at significantly higher expression levels than background genes did. The biological relevance of the predicted expression profiles of trophoblasts (VCT, EVT, and SCT), Endo, Epi, FB, and HB was well validated (**Figure 3A**). However, the predicted expression profiles of granulocytes, T cells, NK, and plasma were found to be more similar to SCT's than to themselves (**Figure 3A**), indicating that the predicted expression profiles of these cells are likely not very useful for further analysis. We further examined the expression levels of the canonical marker genes of the three trophoblast subtypes in these profiles. As the trophoblast stem cell, VCT highly expresses *TOP2A* and *MIK67*, both of which are related to cell proliferation, and the keratin gene *KRT7* is highly expressed in EVT too. The other marker genes of EVT are *HLA-G*, which is involved in the immune tolerance process (Ferreira et al., 2017), and *PRG2* and *DIO2*, both of which are related to the invasion ability of EVT (Windsperger et al., 2017; Adu-Gyamfi et al., 2021). SCT highly expresses *CGA* and *GH1*, which are related to hormone synthesis (Freemark, 2010), and *GDF15*, a classic SCT marker gene, was reported to be associated with PE (Sugulle et al., 2009). Here, these selected marker genes were all highly expressed in their respective predicted cell-type-specific expression profiles (**Figure 3B**). As such, the aforementioned results indicated that the predicted expression profiles of major placental cell types, including Endo, FB, HB, and trophoblasts were worthy of further exploration.

We then focused on the predicted expression profiles of trophoblasts and inspected the activity of several canonical PE-related pathways in between PE and normal controls. As a comparison, we also inspected the activity of these pathways by using the bulk expression profiles. Here, the activity of a pathway was measured by AUCell (Aibar et al., 2017). AUCell sorts all genes in the sample according to their expression and calculates the pathway activity of each sample according to the ranks of the pathway genes. The canonical PE-related pathways inspected here include the epithelial-mesenchymal transition (EMT) hallmark pathway, the hypoxic pathway, and the GO pathway of "Hormone activity."

During the development of trophoblasts (from VCT to EVT and from noninvasive EVT to invasive EVT), the cell undergoes phenotypic changes termed the EMT process in order to gain the invasive ability (Vićovac and Aplin, 1996). It has been well established that the EMT process of trophoblasts was inhibited in PE (Sun et al., 2011). Using the bulk data, however, we did not observe any significant difference in EMT's activity between PE and normal samples (**Figure 3C**). In contrast, in both EVT and VCT, the activity of the EMT pathway was significantly reduced in both EOPE and LOPE though the reduction was not significant in LOPE's EVT (**Figure 3C**), indicating that the invasive ability of EVT and the differentiation of VCT to EVT are likely both inhibited in PE. Not that no EMT-related genes were predicted in SCT.

Placenta hypoxia is one of the most significant clinical manifestations of PE (Soleymanlou et al., 2005). This was clearly shown by using the bulk data: the activity of the hypoxia pathway was significantly upregulated in PE samples (**Figure 3D**). The predicted cell-type-level expression profiles provided more detailed information about hypoxia at the cellular level. In both EOPE and LOPE, the activity of the hypoxia pathway was significantly upregulated in VCT, but not in EVT (**Figure 3D**), reflecting the different pressure of oxygen limitation to different types of trophoblast cells. The significant upregulation of the hypoxia pathway in VCT is probably because VCT is located deeply in the trophoblast layer and is more likely affected by oxygen limitation. Note that there were only a few genes predicted to be associated with the hypoxia pathway in SCT.

It has been reported that the placenta of PE is likely hormonally compensated in response to development deficiency (Tamimi et al., 2003). Here, we observed a significantly higher "Hormone activity" in PE by using the bulk data and further found that the activity was significantly upregulated in SCT, but not in EVT and VCT, by using the predicted cell-type-level expression profiles (**Figure 3E**). Thus, the above results showed that the predicted cell-type-level expression profiles revealed patterns of cell-type-specific gene expression alterations in PE.

As the predicted cell-type-level expression profiles were biologically relevant and provided more details about the altered PE canonical pathways, we explored whether they could better distinguish PE from normal controls than the bulk expression profiles did. For each of the six above-mentioned cell types, we used 80% samples to train an SVM model to distinguish PE from normal samples by using the predicted cell-type-level expression profiles and then tested it using the 20% remaining samples (see **Section 3** for details about the procedures). As a comparison, we also used the bulk expression profiles to develop an SVM model. Overall, it was easier to distinguish EOPE from LOPE; for most cell-type-level SVMs, their performance was comparable to that of bulk-level SVM in EOPE but was superior to LOPE (**Figure 4A,B**). However, even the best SVM in either EOPE or LOPE only achieved an $\mathrm{AUC}_{\mathrm{ROC}}$ less than 0.9, indicating that PE is a heterogeneous and complex disease that may involve multiple subtypes and cannot be easily described by using one model.

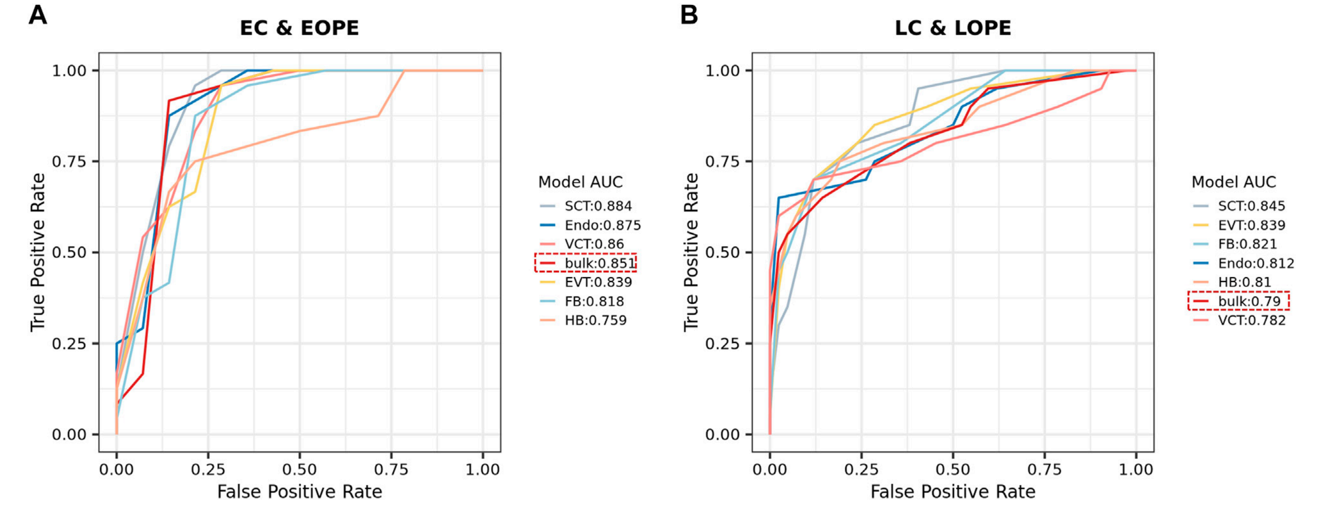


FIGURE 4 | (A,B) ROC of SVM models for distinguishing EOPE **(A)** and LOPE **(B)** from their respective groups of normal samples. SVM models were trained by using either the bulk or the predicted cell-type-level expression profiles.

2.4 Unsupervised Clustering of Predicted Cell-Type-Level Expression Profiles Revealed Clinically Distinct Preeclampsia Subgroups

Although EOPE is generally considered more severe than LOPE, the real situation is usually more complex and severe and nonsevere PE types are actually difficult to distinguish by their subjective clinical indicators (Roberts et al., 2021). The cohort of PE microarray data provided 13 clinical features for a total number of 157 PE samples (EOPE: 80 and LOPE: 77). These features can be divided into two general categories: fetal state-related and maternal state-related. The fetal state-related features include GA, newborn weight z-score, placental weight z-score, umbilical cord diameter, mean umbilical PI, Apgar score (1 min), Apgar score (5 min), and IUGR diagnosis, while the maternal state-related features include maximum systolic bp, maximum diastolic bp, mode proteinuria, mean uterine pi, and maternal BMI. To explore whether PE samples could be classified into subtypes, here for each of the six placental cell types, we conducted unsupervised clustering of PE samples using their predicted expression profiles. Then, we investigated whether the clustering was significantly associated with each of the 13 clinical features.

As a comparison, we first conducted unsupervised clustering of PE samples based on their bulk expression profiles by using negative matrix factorization (NMF) (see **Section 3** for details). We obtained three clusters. The clustering results were found to be significantly associated with not only the definition of EOPE and LOPE but also four fetal state-related features: GA, newborn weight, placental weight, and umbilical cord diameter (**Figure 5A**). Because EOPE and LOPE are defined based on their GA while newborn weight, placental weight, and umbilical cord diameter are also strongly dependent on GA, it is not unexpected that those features were all significantly associated with the clustering results. However, we did not observe any significant maternal state-related clinical features associated with the clustering results.

We next conducted unsupervised clustering of PE samples using the predicted cell-type-level expression profiles of each of the six cell types and investigated their association with clinical features. We found that the clustering results of all six cell types except for SCT were all significantly associated with some clinical features (**Figure 5**). The reason why SCT was not linked to any clinical features was probably that some transcriptional signatures of SCT were misassigned to other cell types, such as NK, granulocytes, and plasma. The clinical features linked to Endo, FB, and VCT were all fetal state-related: Endo was linked to GA, newborn weight z-score, placental weight z-score, and umbilical cord diameter; FB was linked to GA; VCT was linked to mean umbilical PI and umbilical cord diameter (**Figures 5B–D**). Interestingly, the clinical features linked to HB were both fetal state and maternal state-related: newborn weight z-score, placental weight z-score, mode proteinuria, and IUGR diagnosis, while the clinical features linked to EVT were only maternal-related: maximum systolic bp, and maximum diastolic bp (**Figures 5E,F**). HB is an immune cell that promotes trophoblast differentiation and angiogenesis by producing various growth factors and cytokines (Wang and Zhao, 2010). EVT is the primary cell type in the placenta that invades the decidual of the mother during the pregnancy. The reasons why these 2 cell types were linked to maternal state-related features were probably because they had more interaction with maternal cells. In contrast, Endo, FB, and VCT may be more related to the growth of the placenta, that is, more fetus oriented. The predicted cell-type-level expression profiles thus provided more links to clinical features that

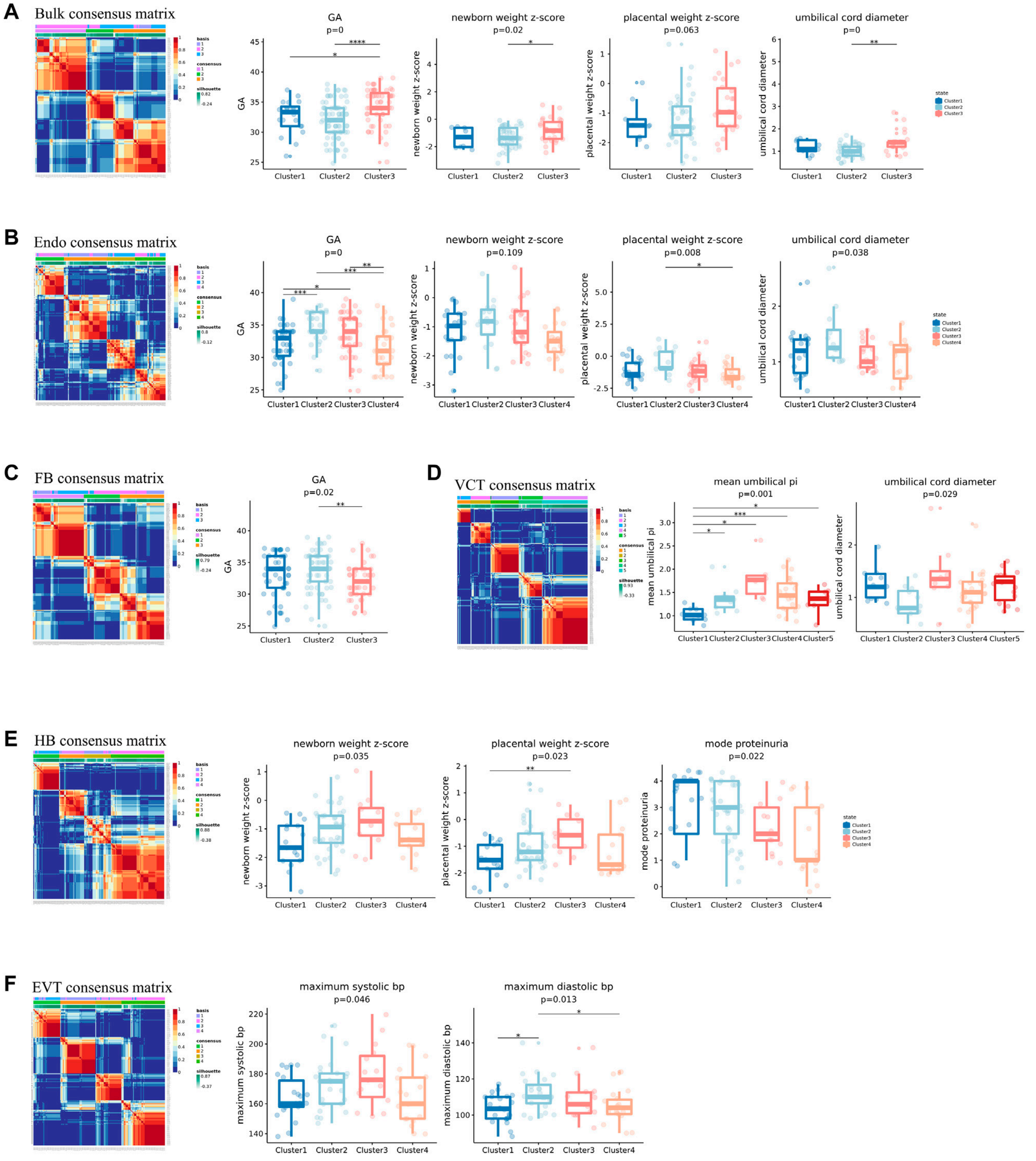


FIGURE 5 | Unsupervised clustering of PE samples using the bulk or predicted cell-type-level expression profiles. The expression profiles used in **(A–F)** were the bulk, the predicted cell-type-level expression profiles of Endo, FB, VCT, HB and EVT, respectively. In each of **(A–F)**, the clinical phenotypes significantly associated with the clustering were shown.

would not be observed by using the bulk expression profiles, especially the maternal state-related features.

Given that the predicted cell-type-level expression profiles of the above five cell types were strongly linked to clinical features, we constructed an integrated expression profile for each sample by combining the predicted expression profiles of the highly variable genes of each cell type and then conducted unsupervised clustering (see **Section 3** for details about constructing the

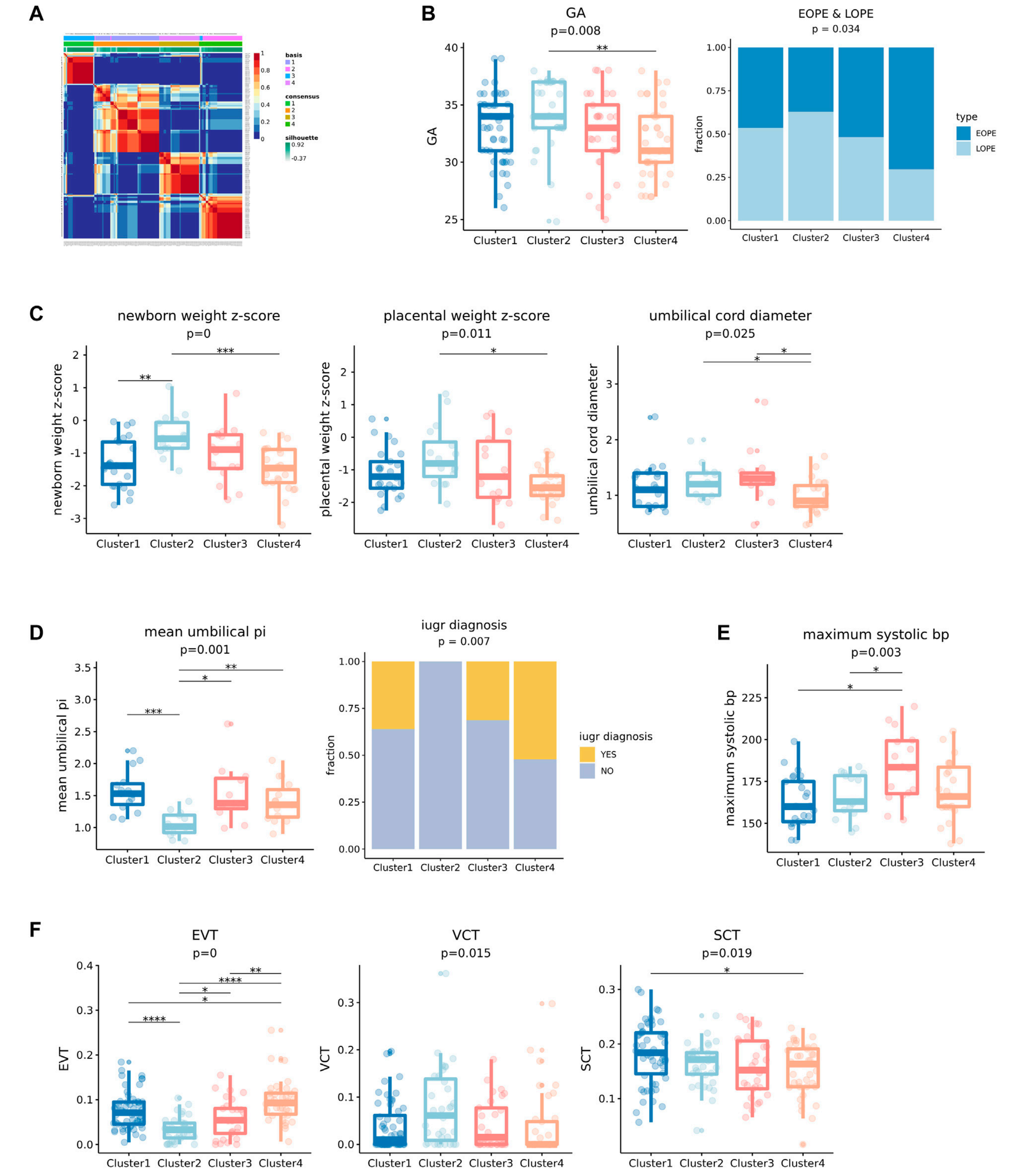


FIGURE 6 | Unsupervised clustering results of integrated five cell types of transcriptional profiles of PE by NMF and significance testing of clinical features. **(A)** Unsupervised clustering result of integrated cell types transcription profiles using NMF, the consensus matrix of NMF output display that four stable clusters can be obtained. **(B)** The four clusters showed significant differences in gestational age (GA), and the fraction of EOPE and LOPE. **(C)** The four clusters showed significant differences in newborn weight z-score, placental weight z-score, and umbilical cord diameter which reflect the state of fetal development. **(D)** The four clusters showed significant differences in mean umbilical PI and the fraction of IUGR in the cluster. A higher mean umbilical PI indicates a greater likelihood of IUGR. **(E)** Four clusters showed significant differences in maximum systolic bp. **(F)** Differences in the proportion of three kinds of trophoblast cells (EVT, VCT, and SCT) in four clusters of PE.

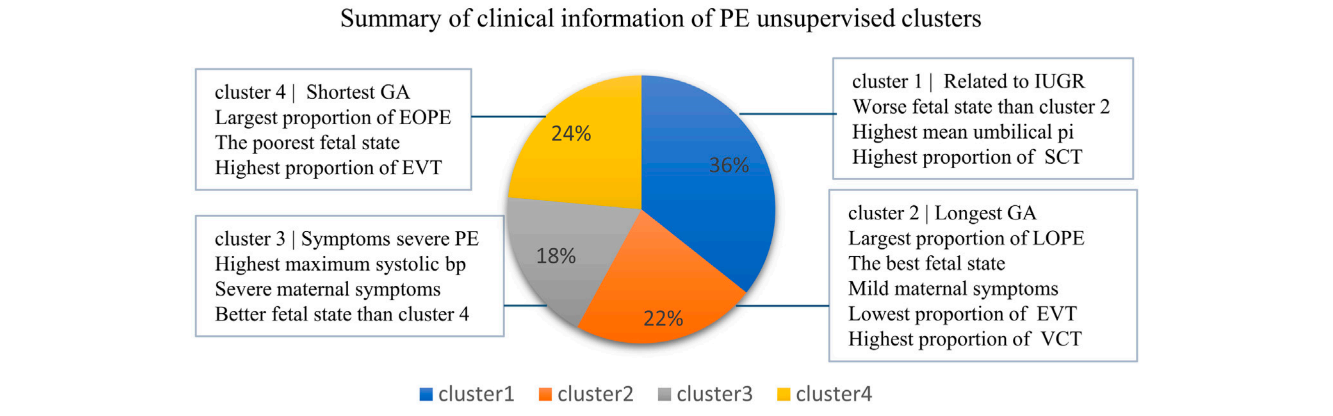


FIGURE 7 | Information overview of four clusters of PE. The sample fractions of the four clusters, respectively, accounted for 36%, 22%, 18%, and 24%.

integrated expression profiles). We obtained four clusters by using NMF (**Figure 6A**) and found that they were significantly associated with seven clinical features of which six were fetal state-related (GA, newborn weight z-score, placental weight z-score, umbilical cord diameter, mean umbilical PI, and IUGR diagnosis) and one was maternal state-related (maximum systolic bp) (**Figures 6B–E**). We compared each of these significant features between the four PE clusters and found that they had distinct clinical features. In general, Clusters 1 and 2 have longer GA, while Clusters 3 and 4 have shorter GA, with Clusters 2 and 4 having the longest and the shortest GA, respectively (**Figure 6B**). Clusters 2 and 4 are also significantly enriched with LOPE and EOPE samples, respectively, while the other two clusters do not have a preference for either EOPE or LOPE (**Figure 6B**). Probably because Clusters 2 and 4 have the longest and shortest GA, they also correspond to the best and the poorest fetal state, respectively (**Figure 6C**). Although Cluster 1's GA is close to Cluster 2's, its fetal state was significantly worse than that of Cluster 2.

For example, Cluster 1 has a significantly higher proportion of intrauterine growth retardation (IUGR), which consists of the higher "mean umbilical PI"—a potential IUGR predictor (Khanduri et al., 2017), compared to Cluster 2 (**Figure 6D**). And its other fetal-related features are also significantly worse than Cluster 2's (**Figure 6C**). Cluster 3's GA is close to Cluster 4's, but it is significantly maternal state-related: it has the highest maximum systolic bp, that is, the most severe state of blood pressure (**Figure 6E**). We also found that the proportions of EVT and VCT were significantly different in these four clusters. For example, the proportion of EVT was the lowest in Cluster 2 which corresponds to the best fetal state, while the proportion of VCT was the highest (**Figure 6F**). Note that when comparing PE samples with normal controls, we observed a significantly increased proportion of EVT and decreased proportion of VCT in PE samples. Therefore, the relative increase or decrease of the proportion of EVT may indicate the severity of PE.

In conclusion, by using the integrated expression profiles, we obtained four clinically distinct PE subtypes that are significantly associated with not only fetal state-related but also maternal state-related clinical features that would not be observed by using the bulk expression profiles (**Figure 7**), highlighting the important value of deconvolution.

3 MATERIALS AND METHODS

3.1 Datasets Used in This Study

A number of PBMC datasets were used for developing the deconvolution pipeline. The bulk PBMC datasets included Finotello's PBMC RNA-seq dataset (Finotello et al., 2019) (GSE107572) and Newman's PBMC microarray dataset (Newman et al., 2015) (GSE65136), and both datasets had known flow-sorting cell-type proportions. The datasets for the reference expression profiles included Hoek's PBMC data (Hoek et al., 2015) with cell-type purified RNA-seq data (GSE64655) and Ding's PBMC dataset (Ding et al., 2020) (https://singlecell.broadinstitute.org/single_cell/study/SCP424) that includes single-cell data produced by 10X, Drop-seq, and inDrops sequencing platforms. The cohort of placenta microarray dataset was built by Leavey et al. (2016) (GSE75010), integrating from 8 placenta microarray studies. It contains 157 samples that had detailed clinical information, including fetal state-related and maternal state-related indicators, and the single-cell placenta reference was generated by Vento-Tormo et al. (2018) (https://www.ebi.ac.uk/arrayexpress/experiments, E-MTAB-6678, E-MTAB-6701). Datasets from GEO were downloaded with accessions above through the website (https://www.ncbi.nlm.nih.gov/geo).

3.2 Procedures for Constructing the Signature Gene Matrix and Description of the Deconvolution Methods Used in the Evaluation

We followed Francisco's recommended strategy on marker gene selection. Given a single-cell reference gene expression matrix, we applied the following parameters to select the marker gene set: logFC ≥ 1 and logCPM ≥ 1. For SecondFC, we determined the relationship between SecondFC and the average correlation between cell types in the signature gene matrix, setting it to no less than 6. Here, logFC means log fold change between the highest expressed cell type and the average expression of other cell types, logCPM means the log average normalized expression level among all cell types, and SecondFC means the average expression fold change of a given marker gene between the highest expressed cell type and the second-highest expressed cell type. When evaluating different deconvolution methods, we ranked the marker genes by logFC and selected a given fraction of top-ranked genes, for example, top 5%, 10%, . . . 100%, and averaged the expression counts of all cells in each cell type to construct the signature gene matrix for the selected marker genes.

We evaluated nine deconvolution methods in this study, among which five bulk reference-based methods and three single-cell reference-based methods were recommended by Francisco et al. (Cobos et al., 2020). The five bulk reference-based methods are nonnegative least squares (NNLS) (https://CRAN.R-project.org/package=nnls), ordinary least squares (OLS) (https://www.R-project.org/), robust linear regression (RLR) (https://www.stats.ox.ac.uk/pub/MASS4/), FARDEEP (https://github.com/YuningHao/FARDEEP), and CIBERSORT (https://cibersort.stanford.edu/), while the three single-cell reference-based methods are DWLS (https://github.com/dtsoucas/DWLS), MuSiC (https://github.com/xuranw/MuSiC), and SCDC (https://github.com/meichendong/SCDC). In addition, we added CIBERSORTx (https://cibersortx.stanford.edu/), which is based on CIBERSORT's improved method of using single-cell data as input.

3.3 The Processing of the Single-Cell Placental Atlas

The single-cell reference expression matrix used for the deconvolution of placental microarray data was constructed from the single-cell placental atlas produced by Vento-Tormo et al. (2018). In order to reduce the problem of collinearity, that is, challenging to the deconvolution algorithm, we merged the subgroups of each of the following cell types in the Vento-Tormo dataset: "DC1" and "DC2" were merged into DC (dendritic) cells, "dNK p," "dNK1," "dNK2," "dNK3," "NK CD16-," and "NK CD16+" were merged into NK (natural killer), "dM1," "dM2," and "dM3" were combined to Mac (macrophage), "Endo (f)," "Endo (m)," and "Endo L" were merged into Endo (endothelial), "Epi1" and "Epi2" were merged into Epi (epithelial), and "fFB1" and "fFB2" were merged into FB (fibroblast). Finally, the single-cell reference expression matrix consisting of a total number of 14 placental cell types was constructed, including eight types of immune cells Hofbauer (HB), NK, T cells, plasma, granulocytes, monocyte (MO), Mac, and DC), three subtypes of trophoblasts (VCT, EVT, and SCT), Epi, Endo, and FB cells. The signature gene matrix was then constructed by applying these cutoffs (logFC ≥ 1, logCPM ≥ 1, and SecondFC ≥ 1.5) and by requiring that each marker gene was expressed in at least 30% of cells of the corresponding cell type.

3.4 The Development of SVM Models to Distinguish PE From Normal Controls

We randomly selected 80% of the samples (training set) to train an SVM model and tested the model using the 20% remaining samples. When training the SVM model, we first identified the differentially expressed genes (DEGs) between PE and normal controls by controlling log CPM >4 using the package of "edgeR" in R. The log-normalized expression profiles of DEGs were then used as the input to train SVM model. For the SVM model, we used svm.SVC classifiers from the scikit-learn library in Python. For the kernel, we chose "linear". For other parameters like degree and gamma, we used the default parameters in the function svm.SVC. For the hyperparameter, C was grid searching between 0 and 2, with 0.2 intervals, and fivefold cross-validation was performed on the training set to find the most appropriate hyperparameter C. The hyperparameter C was determined and then retrained for the whole training set and tested on the test set.

3.5 Procedures of Unsupervised Clustering of Bulk or Predicted Cell-Type-Level Expression Profiles

We first log-normalized raw expression counts and selected highly variable genes by using the "mean.var.plot" method in the Seurat package, with the parameter "mean.cutoff" > 0.5. The "dispersion.cutoff" parameter was tried between 1 and 2.5, with 0.1 intervals, to ensure the stability of unsupervised clustering results. Next, we used the "ScaleData" method to scale the data to maximize the variation between samples. Finally, we used the negative matrix factorization (NMF) to do unsupervised clustering. The input of NMF was the scaled data, and the output of the NMF was the specified k clusters, where k is given artificially. To determine the optimal number of clusters, we iteratively tested k from 2 to 10. In each iteration, we calculated the cophenetic coefficient (CC) of the clusters, which represents the stability of clustering. Ideally, CC remains stable initially when k increases from 2 and then drops quickly when k continues to increase, and the k before the quick drop of CC would be selected. In practice, we would try different values for "dispersion.cutoff" when selecting highly variable genes and selected the one where we could identify the best k.

3.6 Statistical Tests to Inspect the Association of Clinical Features with the Clustering of Preeclampsia Samples

Most of the clinical features are in numerical values. To test the significance of the association of a clinical feature with the clustering of PE samples, when the data type of the clinical feature is numerical, we used analysis of variance (ANOVA) to inspect whether there is any difference in the mean of the clinical features in between clusters, and used *t*-test to check the difference between pairs of clusters, where *p* values were corrected by FDR; when the data type is categorical, we used the Chi-Square test.

4 DISCUSSION

In this study, we aimed to reanalyze a large cohort of PE placental microarray data through deconvolution. For this purpose, we first attempted to develop a practical pipeline by experimenting with the strategies for marker gene selection and several deconvolution methods recommended by Francisco's benchmark study (Cobos et al., 2020). While the selection of marker genes was relatively straightforward, we found that it was not possible to determine which deconvolution method to use by using the metric of PCC_T, the PCC between the predicted expression and true expression of bulk transcripts that can be calculated given estimated cell-type proportions. To have an approximate solution to this problem, we designed several benchmark tests that likely have a similar degree of challenges to the deconvolution of PE placental microarray data and found CIBERSORT performed the best across these tests. CIBERSORT was therefore chosen as the deconvolution method of our study. The successful validation of the biological relevance of the predicted cell-type-level expression profiles of the major placental cell types using their marker genes also confirmed that the deconvolution results by CIBERSORT can be trusted. Based on our experience, the CIBERSORT-based practical pipeline may also well be suited for the deconvolution of other microarray datasets.

In this study, the deconvolution of PE placental microarray data has resulted in several important findings of PE. First, the proportions of EVT and VCT in the placenta are significantly altered in PE, but in different directions, with EVT increasing and VCT decreasing. It has been shown that the differentiation of VCT to EVT and the transition of EVT to gain invasive ability are both inhibited by PE (Sun et al., 2011). Consistently, the activity of the EMT pathway, which plays an important role in these two important development processes (Vićovac and Aplin, 1996), was found to be significantly downregulated in both VCT and EVT in this study. Therefore, the significant increase in EVT and the significant decrease of VCT likely reflect a compensatory enhancement of EVT differentiation and transition in response to the impaired invasive abilities of EVTs. Second, the canonical PE-related pathways showed cell-type-specific alterations in PE. For example, hypoxia was mainly found in VCT, while enhanced hormonal production was found in SCT. Third, placental cell types could be linked to not only fetal state but also maternal state-related clinical features by clustering of predicted cell-type-level expression profiles. In contrast, the clustering of bulk expression profiles could be only linked to fetal state-related clinical features. Although the placenta is a fetus tissue, PE is a disease with significant maternal symptoms, such as high blood pressure. It is therefore of great value that placental cell types, specifically EVT, could be linked to maternal state-related features in our study. Fourth, four clinically distinct clusters of PE samples were identified in this study and likely represent distinct PE subtypes. Clusters 2 and 4 have the longest and the shortest GA and also correspond to the best and the poorest fetal state, respectively. Although Cluster 1 has a similar GA to Cluster 2, it has a significantly much worse fetal state. As for Cluster 3, though it has a similar GA to Cluster 4, it has the most severe maternal state, with the highest blood pressure among the four clusters.

The discovery of clinically distinct clusters by this study is of great value to the field of PE. For example, a new diagnostic model can be developed based on the classification of these clinically distinct clusters, such that PE patients can be assigned into different groups and different treatment plans can be applied. New therapeutic drugs targeting the most severe PE may also be developed by selecting drug target genes from the marker genes from the PE cluster with the most severe outcomes. Moreover, there is a rich trove of bulk RNA-seq or microarray data in the public domain, with many having disease-related clinical information. The fact that the deconvolution of PE placental microarray data led to several new findings on the disease strongly suggests that similar deconvolution studies should be conducted to reanalyze disease-related bulk data to generate new insights into the molecular mechanisms of diseases.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75010.

AUTHOR CONTRIBUTIONS

WT conceived of the study. TY and QL designed the experiments and performed the analysis. TY, QL, and WT drafted the manuscript.

FUNDING

This work was supported by the National Key Research and Development Program of China (2021YFC2301500) and the National Natural Science Foundation of China (32170667, 31871325).

REFERENCES

Adu-Gyamfi, E. A., Lamptey, J., Chen, X.-M., Li, F.-F., Li, C., Ruan, L.-L., et al. (2021). Iodothyronine Deiodinase 2 (DiO2) Regulates Trophoblast Cell Line Cycle, Invasion and Apoptosis; and its Downregulation Is Associated with Early Recurrent Miscarriage. *Placenta* 111, 54–68. doi:10.1016/j.placenta.2021.06.004

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 (11), 1083–1086. doi:10.1038/nmeth.4463

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

Bokslag, A., van Weissenbruch, M., Mol, B. W., and de Groot, C. J. M. (2016). Preeclampsia; Short and Long-Term Consequences for Mother and Neonate. *Early Hum. Dev.* 102, 47–50. doi:10.1016/j.earlhumdev.2016.09.007

Chambers, J., Hastie, T., and Pregibon, D. (1990). "Statistical Models in S," in *Compstat* (Heidelberg: Physica-Verlag HD).

Cobos, F. A., José, A-H., Joseph, E. P., Pieter, M., and Katleen De, P. (2020). Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data. *Nat. Commun.* 11 (1). doi:10.1038/s41467-020-19015-1

Crosley, E. J., Elliot, M. G., Christians, J. K., and Crespi, B. J. (2013). Placental Invasion, Preeclampsia Risk and Adaptive Molecular Evolution at the Origin of the Great Apes: Evidence from Genome-wide Analyses. *Placenta* 34 (2), 127–132. doi:10.1016/j.placenta.2012.12.001

Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., et al. (2020). Systematic Comparison of Single-Cell and Single-Nucleus RNA-Sequencing Methods. *Nat. Biotechnol.* 38 (6), 737–746. doi:10.1038/s41587-020-0465-8

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., et al. (2021). SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References. *Briefings Bioinforma.* 22 (1), 416–427. doi:10.1093/bib/bbz166

Ferreira, L. M. R., Meissner, T. B., Tilburgs, T., and Strominger, J. L. (2017). HLA-G: At the Interface of Maternal-Fetal Tolerance. *Trends Immunol.* 38 (4), 272–286. doi:10.1016/j.it.2017.01.009

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., et al. (2019). Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-Seq Data. *Genome Med.* 11 (1), 34. doi:10.1186/s13073-019-0638-6

Freemark, M. (2010). Placental Hormones and the Control of Fetal Growth. *J. Clin. Endocrinol. Metab.* 95, 2054–2057. doi:10.1210/jc.2010-0517

Gaujoux, R., and Seoighe, C. (2012). Semi-supervised Nonnegative Matrix Factorization for Gene Expression Deconvolution: a Case Study. *Infect. Genet. Evol.* 12 (5), 913–921. doi:10.1016/j.meegid.2011.08.014

Hao, Y., Yan, M., Heath, B. R., Lei, Y. L., and Xie, Y. (2019). Fast and Robust Deconvolution of Tumor Infiltrating Lymphocyte from Expression Profiles Using Least Trimmed Squares. *PLoS Comput. Biol.* 15 (5), e1006976. doi:10.1371/journal.pcbi.1006976

Hoek, K. L., Samir, P., Howard, L. M., Niu, X., Prasad, N., Galassie, A., et al. (2015). A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination. *PLoS One* 10 (2), e0118528. doi:10.1371/journal.pone.0118528

Horii, M., Morey, R., Bui, T., Touma, O., Nelson, K. K., Cho, H.-Y., et al. (2021). Modeling Preeclampsia Using Human Induced Pluripotent Stem Cells. *Sci. Rep.* 11 (1), 5877. doi:10.1038/s41598-021-85230-5

Jin, H. J., and Liu, Z. D. (2021). A Benchmark for RNA-Seq Deconvolution Analysis under Dynamic Testing Environments. *Genome Biol.* 22 (1). doi:10.1186/s13059-021-02290-6

Katharine, M., Mullen, K. M., Ivo, H. M., and van Stokkum, K. M. (2012). nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). *R package version 1.4.* Available at: https://CRAN.R-project.org/package=nnls

Khanduri, S., Chhabra, S., Yadav, S., Sabharwal, T., Chaudhary, M., Usmani, T., et al. (2017). Role of Color Doppler Flowmetry in Prediction of Intrauterine Growth Retardation in High-Risk Pregnancy. *Cureus* 9 (11), e1827. doi:10.7759/cureus.1827

Leavey, K., Benton, S. J., Grynspan, D., Kingdom, J. C., Bainbridge, S. A., and Cox, B. J. (2016). Unsupervised Placental Gene Expression Profiling Identifies Clinically Relevant Subclasses of Human Preeclampsia. *Hypertension* 68 (1), 137–147. doi:10.1161/hypertensionaha.116.07293

Nadel, B. B., Oliva, M., Shou, B. L., Mitchell, K., Ma, F., Montoya, D. J., et al. (2021). Systematic Evaluation of Transcriptomics-Based Deconvolution Methods and References Using Thousands of Clinical Samples. *Brief. Bioinform* 22 (6). doi:10.1093/bib/bbab265

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry. *Nat. Biotechnol.* 37 (7), 773–782. doi:10.1038/s41587-019-0114-2

Roberts, J. M., Rich-Edwards, J. W., McElrath, T. F., Garmire, L., and Myatt, L. (2021). Subtypes of Preeclampsia: Recognition and Determining Clinical Usefulness. *Hypertension* 77 (5), 1430–1441. doi:10.1161/hypertensionaha.120.14781

Robineau-Charette, P., Grynspan, D., Benton, S. J., Gaudet, J., Cox, B. J., Vanderhyden, B. C., et al. (2020). Fibrinogen-Like Protein 2-Associated Transcriptional and Histopathological Features of Immunological Preeclampsia. *Hypertension* 76 (3), 910–921. doi:10.1161/hypertensionaha.120.14807

Soleymanlou, N., Jurisica, I., Nevo, O., Ietta, F., Zhang, X., Zamudio, S., et al. (2005). Molecular Evidence of Placental Hypoxia in Preeclampsia. *J. Clin. Endocrinol. Metabolism* 90 (7), 4299–4308. doi:10.1210/jc.2005-0078

Steen, C. B., Liu, C. L., Alizadeh, A. A., and Newman, A. M. (2020). "Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx," in *Stem Cell Transcriptional Networks: Methods and Protocols*. Editor B. L. Kidder (New York, NY: Springer US), 135–157. doi:10.1007/978-1-0716-0301-7_7

Sugulle, M., Dechend, R., Herse, F., Weedon-Fekjaer, M. S., Johnsen, G. M., Brosnihan, K. B., et al. (2009). Circulating and Placental Growth-Differentiation Factor 15 in Preeclampsia and in Pregnancy Complicated by Diabetes Mellitus. *Hypertension* 54 (1), 106–112. doi:10.1161/hypertensionaha.109.130583

Sun, Y.-Y., Lu, M., Xi, X.-W., Qiao, Q.-Q., Chen, L.-L., Xu, X.-M., et al. (2011). Regulation of Epithelial-Mesenchymal Transition by Homeobox GeneDLX4in JEG-3 Trophoblast Cells: A Role in Preeclampsia. *Reprod. Sci.* 18 (11), 1138–1145. doi:10.1177/1933719111408112

Suryawanshi, H., Morozov, P., Straus, A., Sahasrabudhe, N., Max, K. E. A., Garzia, A., et al. (2018). A Single-Cell Survey of the Human First-Trimester Placenta and Decidua. *Sci. Adv.* 4 (10), eaau4788. doi:10.1126/sciadv.aau4788

Tamimi, R., Lagiou, P., Vatten, L. J., Mucci, L., Trichopoulos, D., Hellerstein, S., et al. (2003). Pregnancy Hormones, Pre-eclampsia, and Implications for Breast Cancer Risk in the Offspring. *Cancer Epidemiol. Biomarkers Prev.* 12 (7), 647–650.

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G. C. (2019). Accurate Estimation of Cell-type Composition from Gene Expression Data. *Nat. Commun.* 10, 2975. doi:10.1038/s41467-019-10802-z

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition*. New York: Springer.

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., et al. (2018). Single-cell Reconstruction of the Early Maternal-Fetal Interface in Humans. *Nature* 563 (7731), 347–353. doi:10.1038/s41586-018-0698-6

Vićovac, L., and Aplin, J. D. (1996). Epithelial-mesenchymal Transition during Trophoblast Differentiation. *Acta Anat. (Basel)* 156 (3), 202–216. doi:10.1159/000147847

Von Dadelszen, P., Magee, L. A., and Roberts, J. M. (2003). Subclassification of Preeclampsia. *Hypertens. pregnancy* 22 (2), 143–148. doi:10.1081/prg-120021060

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, 380. doi:10.1038/s41467-018-08023-x

Wang, Y., and Zhao, S. (2010). "Integrated Systems Physiology: from Molecules to Function to Disease," in *Vascular Biology of the Placenta* (San Rafael (CA): Morgan & Claypool Life Sciences Copyright © 2010 by Morgan & Claypool Life Sciences).

Windsperger, K., Dekan, S., Pils, S., Golletz, C., Kunihs, V., Fiala, C., et al. (2017). Extravillous Trophoblast Invasion of Venous as Well as Lymphatic Vessels Is Altered in Idiopathic, Recurrent, Spontaneous Abortions. *Hum. Reprod.* 32 (6), 1208–1217. doi:10.1093/humrep/dex058

Xu, H., Yin, X., Yanan, S., Rong, G., Dan, L., Xuanxuan, L., et al. (2021). Integrated Analysis of Multiple Microarray Studies to Identify Potential Pathogenic Gene Modules in Preeclampsia. *Exp. Mol. Pathology* 120, 104631. doi:10.1016/j.yexmp.2021.104631

Yadama, A. P., Maiorino, E., Carey, V. J., McElrath, T. F., Litonjua, A. A., Loscalzo, J., et al. (2020). Early-pregnancy Transcriptome Signatures of Preeclampsia: from Peripheral Blood to Placenta. *Sci. Rep.* 10 (1), 17029. doi:10.1038/s41598-020-74100-1

Zhang, S., Zhang, E., Long, J., Hu, Z., Peng, J., Liu, L., et al. (2019). Immune Infiltration in Renal Cell Carcinoma. *Cancer Sci.* 110 (5), 1564–1572. doi:10.1111/cas.13996

Zhang, T., Qianqian, B., Yanchun, C., Xiaolin, W., Shaowei, Y., Shunhua, L., et al. (2021). Dissecting Human Trophoblast Cell Transcriptional Heterogeneity in Preeclampsia Using Single-Cell RNA Sequencing. *Mol. Genet. Genomic Med.* 9 (8). doi:10.1002/mgg3.1730

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.

TYPE Original Research
PUBLISHED 25 August 2022
DOI 10.3389/fbioe.2022.892613

The microbiome of lower respiratory tract and tumor tissue in lung cancer manifested as radiological ground-glass opacity

Zhigang Wu[1†], Jie Tang[1†], Runzhou Zhuang[1], Di Meng[1], Lichen Zhang[1], Chen Gu[1], Xiao Teng[1], Ziyue Zhu[1], Jiacong Liu[1], Jinghua Pang[2], Jian Hu[1]* and Xiayi Lv[1]*

[1]Department of Thoracic Surgery, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China, [2]Department of Thoracic Surgery, Fenghua People's Hospital, Ningbo, China

Recent studies have confirmed the existence of microbiota in the lungs. The relationship between lung ground-glass opacity (GGO) and microbiota in the lung microenvironment is not clear. In this study, we investigated the microbial composition and diversity in bronchoalveolar lavage fluid (BALF) of diseased lung segments and paired contralateral healthy lung segments from 11 GGO patients. Furthermore, lung GGO and paired normal tissues of 26 GGO patients were explored whether there are microbial characteristics related to GGO. Compared with the control group, the community richness of GGO tissue and BALF of GGO lung segment (α-diversity) and overall microbiome difference (β-diversity) had no significant difference. The microbiome composition of BALF of GGO segments is distinct from that of paired healthy lung segments [genus (*Rothia*), order (*Lachnospiraceae*), family (*Lachnospiraceae*), genus (*Lachnospiraceae_NK4A136_group*, *Faecalibacterium*), and species (*Faecalibacterium prausnitzii*, *Bacteroides uniforms*)]. GGO tissue and adjacent lung tissue had more significant differences at the levels of class, order, family, genus, and species level, and most of them are enriched in normal lung tissue. The area under the curve (AUC) using 10 genera-based biomarkers to predict GGO was 91.05% (95% CI: 81.93–100%). In conclusion, this study demonstrates there are significant differences in the lower respiratory tract and lung microbiome between GGO and the non-malignant control group through the BALF and lung tissues. Furthermore, some potential bacterial biomarkers showed good performance to predict GGO.

KEYWORDS
microbiome, 16S rRNA sequencing, ground-glass opacity, lung cancer, biomarker

OPEN ACCESS

EDITED BY
Zhaoyuan Fang,
Zhejiang University, China

REVIEWED BY
Dinesh Devadoss,
Florida International University, United States
Lianying Jiao,
Xi'an Jiaotong University, China

*CORRESPONDENCE
Jian Hu,
dr_hujian@zju.edu.cn
Xiayi Lv,
lyuxiayi@zju.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

RECEIVED 09 March 2022
ACCEPTED 30 June 2022
PUBLISHED 25 August 2022

CITATION
Wu Z, Tang J, Zhuang R, Meng D, Zhang L, Gu C, Teng X, Zhu Z, Liu J, Pang J, Hu J and Lv X (2022), The microbiome of lower respiratory tract and tumor tissue in lung cancer manifested as radiological ground-glass opacity.
Front. Bioeng. Biotechnol. 10:892613.
doi: 10.3389/fbioe.2022.892613

Introduction

Low-dose computed tomography (LDCT) is widely used as the main method of lung cancer screening project wordwide, and an increasing number of lung ground-glass opacity (GGO) are found (Aberle et al., 2011). Pulmonary nodules are round shadows with a diameter of less than 3 cm in chest CT images. Among pulmonary nodules, GGO are defined as lesions with higher opacity than normal lung tissue, but lower than the consolidated bronchovascular edge (MacMahon et al., 2017). Although GGO is a non-specific radiologic manifestation, persistent and long-term stable GGO is generally considered to be malignant and still considered to be an inert and progressed slowly subtype of lung adenocarcinoma (Chang et al., 2013). Therefore, the causes of GGO have attracted the attention of clinicians and researchers. GGO usually does not have driver gene mutations, which usually occurred in lung adenocarcinomas, such as EGFR and ALK (Ren et al., 2019a). Benign lesions including infectious diseases such as COVID-19 can also be radiographed as GGO. Pathologically, GGO can be caused by interstitial thickening with inflammation, edema, fibrosis, and tumor proliferation (Fan et al., 2012). Meanwhile, epidemiological studies have also suggested that there is a close relationship between chronic infection, inflammation, and lung cancer (Gomes et al., 2014). Therefore, the microbiome may play an important role in the occurrence of early adenocarcinoma characterized by GGO.

At present, the most research on microbiome and diseases is the correlation between intestinal microbiome and some metabolic diseases or gastrointestinal cancer. However, with the development of high-throughput next-generation sequencing (NGS), the entire spectrum of the human microbiome has been surveyed; recent studies show that in addition to intestinal microbiome, symbiotic microbiome also exists in other locations of the human body. In the past, it was considered that the lung is a sterile space, but recent studies have suggested that the lower respiratory tract is also full of various bacterial communities, which is very important in maintaining the stability of the internal environment and can cause respiratory diseases such as asthma, COPD, and lung cancer (Hilty et al., 2010; Mao et al., 2018; Maddi et al., 2019; Ramírez-Labrada et al., 2020). Compared with gastrointestinal cancer, there are few studies on the correlation between microbiome and lung cancer. Epidemiological studies have shown the correlation between repeated exposure to antibiotics and increased risk of lung cancer (Boursi et al., 2015), but the effect of lung microbiome on lung cancer is still unknown.

Some studies have confirmed that there are some unique microbiota in BALF, sputum, saliva, or lung tissue of patients with lung cancer (Lee et al., 2016; Yu et al., 2016; Cameron et al., 2017; Liu et al., 2018a; Tsay et al., 2018; Peters et al., 2019; Zhang et al., 2019; Mao et al., 2020), and these studies have not only found similar but also contradictory microbiota prevalent in patients with lung cancer. However, previous studies have mostly compared the microbial composition of bronchoalveolar lavage fluid from lung cancer patients and healthy people, or tumor tissue and normal lung tissue from typical lung cancer patients. Few studies explored the microbiome composition of GGO lesions. In this study, we screened BALF from 11 patients with GGO, fresh frozen GGO lung tissue, and paired adjacent lung tissue from 26 patients. Furthermore, the microbial diversity of lower respiratory tract and lung tissue of patients with GGO and the identified characteristic microbiome were revealed, which also provides a new idea for the occurrence and treatment of GGO.

Materials and methods

Patient enrollment and sample collection

The enrolled patients were from patients who underwent radical resection of lung cancer in the First Affiliated Hospital of Medical College of Zhejiang University from September 2019 to September 2021. Twenty six lung tumor specimens and paired normal lung tissues were collected, and bronchoalveolar lavage fluid of 11 diseased lung segments and paired contralateral healthy lung segments were collected. The included patients did not use antibiotics or adjuvant therapy 3 months before operation; HRCT showed pulmonary ground-glass nodules; lung cancer was diagnosed by pathology; and no previous history of other cancers. Bronchoalveolar lavage fluid (15ml) from the diseased lung segment and the contralateral healthy lung segment was centrifuged and enriched and put into liquid nitrogen. The tumor tissue was removed under sterile conditions and immediately put into liquid nitrogen, and then transferred to the −80° refrigerator for preservation until DNA extraction. While collecting tumor tissue, collect adjacent normal lung tissue more than 5 cm away from tumor lesion to avoid local influence of tumor. At the same time, a blank control tube is designed to run through the whole sample collection process, and then delivered in dry ice container to Novogene Inc. (Beijing, China) for 16S rRNA gene sequencing.

DNA extraction

Total genome DNA from BALF and lung tissue samples was extracted using the CTAB method. DNA concentration and purity were monitored on 1% agarose gels. According to the concentration, DNA was diluted to 1 ng/μl using sterile water.

16S rRNA gene sequencing

16S rRNA genes of distinct regions (16S V4/16S V3/16S V3-V4/16S V4-V5) were amplified using specific primer with the barcode. All PCR reactions were carried out with 15 μl of Phusion® High-Fidelity PCR Master Mix (New England Biolabs). Mix same volume of 1X loading buffer (contained SYBR Green) with PCR products and operate electrophoresis on 2% agarose gel for detection. PCR products were mixed in equidensity ratios. Then, mixture PCR products were purified with the Qiagen Gel Extraction Kit (Qiagen, Germany). Sequencing libraries were generated using TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, United States), following manufacturer's recommendations, and index codes were added. The library quality was assessed on the Qubit@ 2.0 Fluorometer (Thermo Scientific) and Agilent Bioanalyzer 2100 system. At last, the library was sequenced on an Illumina NovaSeq platform and 250 bp paired-end reads were generated.

Data analysis

Sequences with ≥97% similarity were assigned to the same OTUs. Representative sequence for each OTU was screened for further annotation. OTUs abundance information was normalized using a standard of the sequence number, corresponding to the sample with the least sequences. Subsequent analysis of alpha diversity and beta diversity were all performed based on this output normalized data. Alpha diversity is applied in analyzing complexity of species diversity for a sample through two indices, including Shannon and Simpson. All this indices in our samples were calculated with QIIME (Version 1.7.0) and displayed with R software (Version 2.15.3). Beta diversity analysis was used to evaluate differences of samples in species complexity; beta diversity on both weighted and unweighted UniFrac was calculated by QIIME software (Version 1.9.1). PERMANOVA was used to test the statistical significance of diversity differences between groups. The linear discriminant analysis (LDA) score by LEfSe (LDA effect size) was used to estimate taxa features with significant differential abundance. The random forest model was performed to estimate the importance of each differential genus and test predictive power based on the area under the receiver operating characteristic curve (ROC).

Results

Patient characteristics

A total of 37 patients with ground-glass nodules were included in the study. All patients had no other lung comorbidities. They were confirmed as lung cancer by pathology. Among all patients, 11 patients underwent bronchoscopy, and bronchoalveolar lavage fluid was collected before operation, and 26 patients underwent radical resection of lung cancer and collected surgical specimens. The clinical characteristics of the two groups of patients are shown in Table 1.

Lower respiratory tract microbiota in lung segment with ground-glass opacity and contralateral normal lung segment

Splicing and quality control were performed to obtain effective tags for subsequent analysis through the Illumina NovaSeq sequencing platform. An average of 87,624 tags was measured per sample, and an average of 70,471 valid data was obtained after quality control. The effective rate of quality control was 80%. The operational taxonomic units (OTUs) were clustered with 97% identity, and a total of 4,272 OTUs were obtained with 3,685 OTUs in BALF of a lung segment with GGO and 3,365 in BALF of a contralateral normal lung segment (Figure 1A), and the sequence of OTUs was annotated finally. The richness and diversity of microbial community (α-diversity) in BALF samples of the lung segment with GGO and contralateral normal lung segment were measured by Chao1 index, Shannon index, and Simpson index had no significant difference (Figure 1C). PERMANOVA analysis based on the Bray–Curtis dissimilarity (Figure 1D), unweighted, and weighted UniFrac boxplot (Supplementary Figure S1) revealed that there were no significant differences in the overall microbiota (β-diversity) between two groups of BALF.

According to the relative abundance of the microbiota in the BALF samples of the two groups, classification and analysis were based on the phylum, class, order, family, genus, and species levels (Supplementary Figure S2). At the phylum level, the most abundant compositions were *Bacteroidota*, *Proteobacteria*, *Firmicutes*, *Fusobacteriota*, and *Actinobacteria* in both BALF of the lung segment with GGO and contralateral normal lung segment (Figure 1B). However, there was no significant difference in the phylum level of the main flora between the two groups. In addition, at the genus level, *Rothia* is more enriched in BALF of the normal lung segment ($p < 0.05$) (Figure 2). Furthermore, the relative abundance of microbiota at order (*Lachnospiraceae*), family (*Lachnospiraceae*), genus (*Lachnospiraceae_NK4A136_group, Faecalibacterium*), and species (*Faecalibacterium prausnitzii*, *Bacteroides uniforms*) level is increased significantly in BALF of the lung segment with GGO ($p < 0.05$) (Figure 2).

The composition and diversity of lung microbiota in lung ground-glass opacity and paired adjacent normal tissue

Based on the Illumina NovaSeq sequencing platform, lung tissue samples were sequenced and analyzed similar to BALF samples to obtain OTUs for subsequent analysis. GGO tumor tissue and paired adjacent normal lung tissue had the same total of 4,491 OTUs, which is much more than BALF samples (Figure 3A). The main phyla in the microbiome of GGO

TABLE 1 Baseline clinical characteristics of the study cohort.

Clinical characteristic	BALF group (n = 11)	GGO group (n = 26)
Age (years; mean ± SD)	51.81 ± 9.06	51.54 ± 9.84
Sex (female)	9 (81.82%)	20 (76.92%)
Smoking (yes)	1 (9.09%)	5 (23.08%)
Multiple (yes)	2 (18.18%)	7 (26.92%)
Lesion location		
Upper left	3 (27.27%)	6 (23.08%)
Lower left	2 (18.18%)	4 (15.38%)
Upper right	3 (27.27%)	10 (38.46)
Middle-lower right	3 (27.27%)	6 (23.08)
Surgery type		
Wedge resection	6 (54.55%)	14 (53.85%)
Segmentectomy	5 (45.45%)	9 (34.62%)
Lobectomy	0	3 (11.53%)
Tumor diameter (cm; mean ± SD)	0.82 ± 0.15	0.91 ± 0.23
Histology		
AIS	0	2 (7.69%)
MIA	7 (63.63%)	16 (61.54%)
IAC	4 (36.37%)	8 (30.77%)

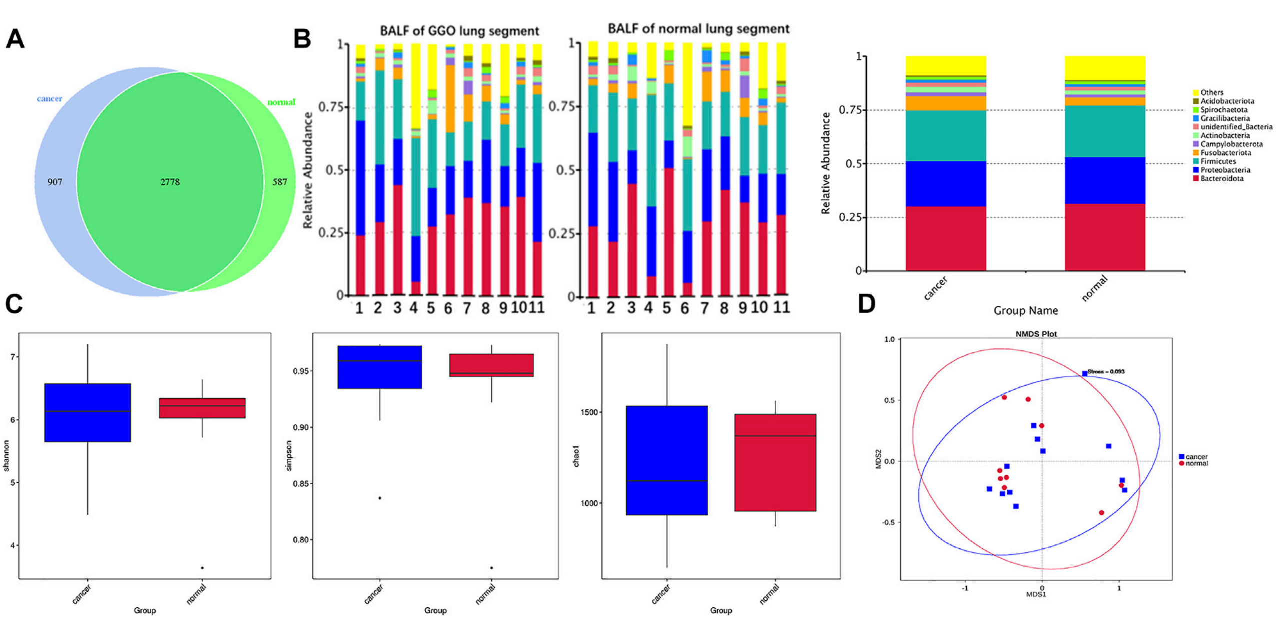


FIGURE 1
Microbial composition and diversity in BALF of the lung segment with GGO and contralateral normal lung segment. **(A)** Operational taxonomic units (OTUs) between GGO and normal BALF groups. **(B)** Bar plot presents the relative abundance of microbial phyla in each sample and groups. **(C)** Shannon, Simpson, and Chao1 index of GGO and normal BALF groups ($p > 0.05$). **(D)** Non-metric multidimensional scaling (NMDS) plot visualizes the overall microbiome dissimilarity measured by Bray–Curtis dissimilarities.

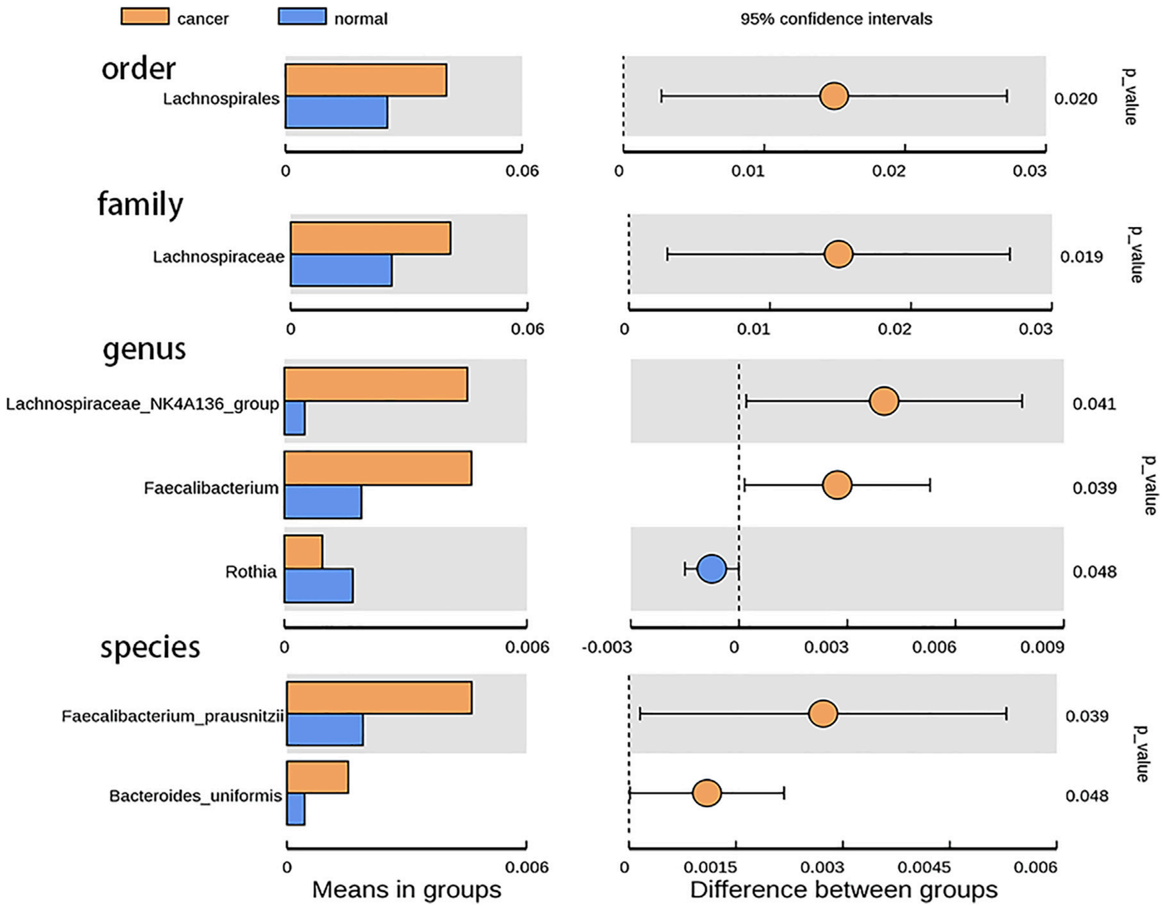


FIGURE 2
Bar plot presents the microbiota with significant differential relative abundance on the phylum, class, order, family, genus, and species levels between BALF of lung segment with GGO and contralateral normal lung segment.

tissues and adjacent non-malignant tissues include *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes*, and the most abundant genera were *Ralstonia*, *Herbaspirillum*, and *Sphingomonas* (Figure 3B). In addition, there are significant differences in *Proteobacteria* between tumor tissues and normal tissues in main phyla (Figure 4). However, α-diversity which was estimated by Chao1 index, Shannon index, and Simpson index and PERMANOVA analysis (β-diversity) based on Bray–Curtis dissimilarity (Figures 3C,D), unweighted, and weighted UniFrac boxplot were of no significant difference between GGO tissues and adjacent tissues, which was the same as BALF samples. GGO tissue and adjacent lung tissue had significant differences in the composition of flora at the levels of class, order, family, genus, and species as shown in Figure 4, and interestingly most of them are enriched in normal lung tissue.

Potential biomarkers for ground-glass opacity based on bacterial taxa feature

The receiver operating characteristic (ROC) analysis was performed to evaluate the diagnostic ability of potential biomarkers in GGO based on the 10 different genera of GGO tissue and adjacent normal lung tissue, and the calculated area under the curve (AUC) represented the diagnostic performance of each biomarker. The AUC produced by 10 difference genera was 91.05% (95% CI: 81.93–100%) (Figure 5A), which were proven to be effective in distinguishing GGO and paired adjacent normal tissue. The importance ranking of the 10 difference genera included in the random forest analysis was demonstrated by mean decrease accuracy (Figure 5B) and mean decrease Gini (Figure 5C).

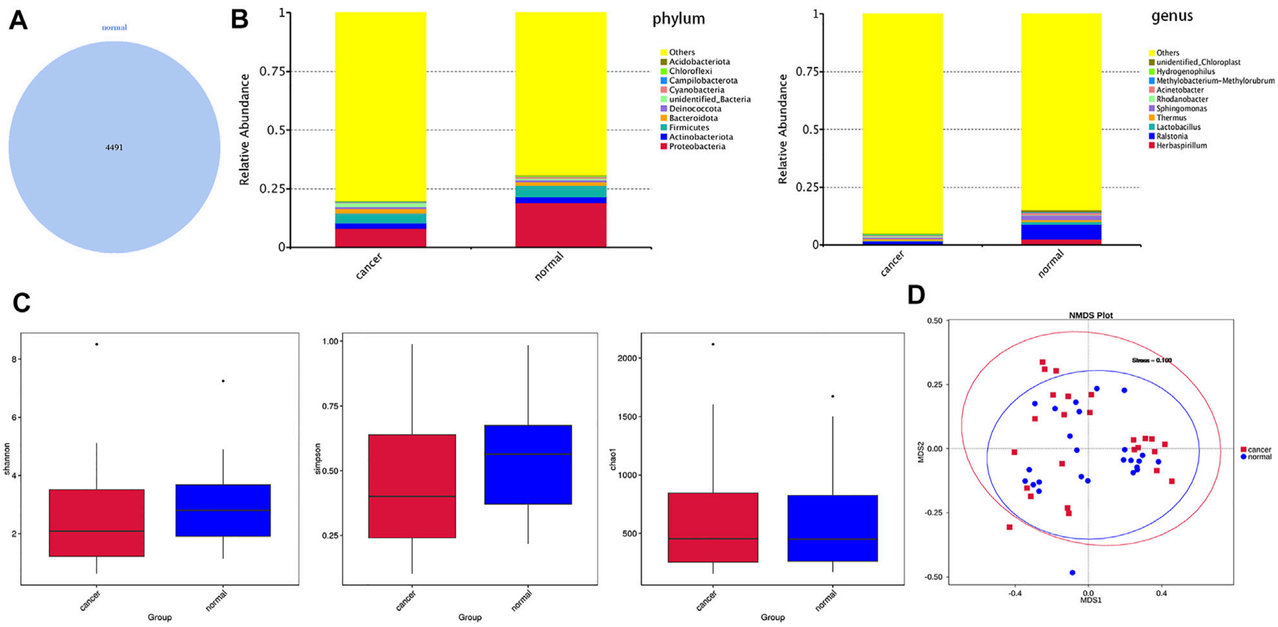


FIGURE 3

Microbial composition and diversity in lung GGO and paired adjacent normal tissues. **(A)** Operational taxonomic units (OTUs) between GGO and normal tissues. **(B)** Bar plot presents the relative abundance of microbial phyla and genera in GGO and normal lung tissues. **(C)** Shannon, Simpson, and Chao1 index of GGO and normal groups ($p > 0.05$). **(D)** Non-metric multidimensional scaling (NMDS) plot visualizes the overall microbiome dissimilarity measured by Bray–Curtis dissimilarities.

Discussion

The past view was that the lungs of healthy people were sterile. However, with the development of high-throughput NGS, several studies recently confirmed the existence of microbiota in healthy lungs (Dickson et al., 2016), which overturned the past cognition, and the lung microbiota was associated with human health and disease status and played an important role in cancer progressing (Maddi et al., 2019). In this study, we confirmed that there was obvious microbiota in the BALF and tissue samples of patients with GGO, which provides the research direction and clue of tumor microbiome for the occurrence of GGO.

Because the bacteria content of healthy lung is very small, the external pollution in the process of sample collection and experiment has a great impact on the results (Salter et al., 2014). It is very important to set up a negative control in the study of lung microbiome. In our study, we collected BALF and lung tissue removed by aseptic surgery. In order to avoid pollution in the process of sampling and DNA extraction, we designed a negative control respectively. The results showed that the DNA concentration of the negative control was very low and could not be amplified by PCR, which ruled out the influence of external pollution on the results.

In the current study of lung microbiome, most of them are studied through BALF and brush samples. Because bronchoscopy needs to enter the lower respiratory tract through the upper respiratory tract, there is a risk of sample contamination. However, studies have shown that the microbiota in BALF obtained by bronchoscopy is not affected (Dickson et al., 2016). Therefore, BALF is feasible as a research method of lower respiratory tract microbiome.

Some studies have confirmed that in chronic lung diseases, the flora structure of lower respiratory tract will change, such as COPD (Mammen and Sethi, 2016) and bronchiectasis (Budden et al., 2019). Differences in the overall structure of lung microbiome composition between lung cancer and non-malignant diseases were observed, which was consistent with the results of Liu et al. (2018b); Tsay et al. (2018), indicating that there were significant differences in the composition of pulmonary microbial communities between the two groups. In our study, we found that there were differences in the overall structure of microbial communities between the two groups by NMDS analysis, which suggested that there were significant differences in the microbial composition of the lower respiratory tract between the lung segment with GGO and the contralateral normal lung segment. The results of a diversity analysis showed that there was no significant difference in the richness and diversity of microbiota between the BALF samples of the diseased and normal lung segment, which was similar with the conclusions of Jin's study on BALF microbiome in patients with lung cancer and healthy patients (Jin et al., 2019), indicated that the microbiome composition of the lower respiratory tract is

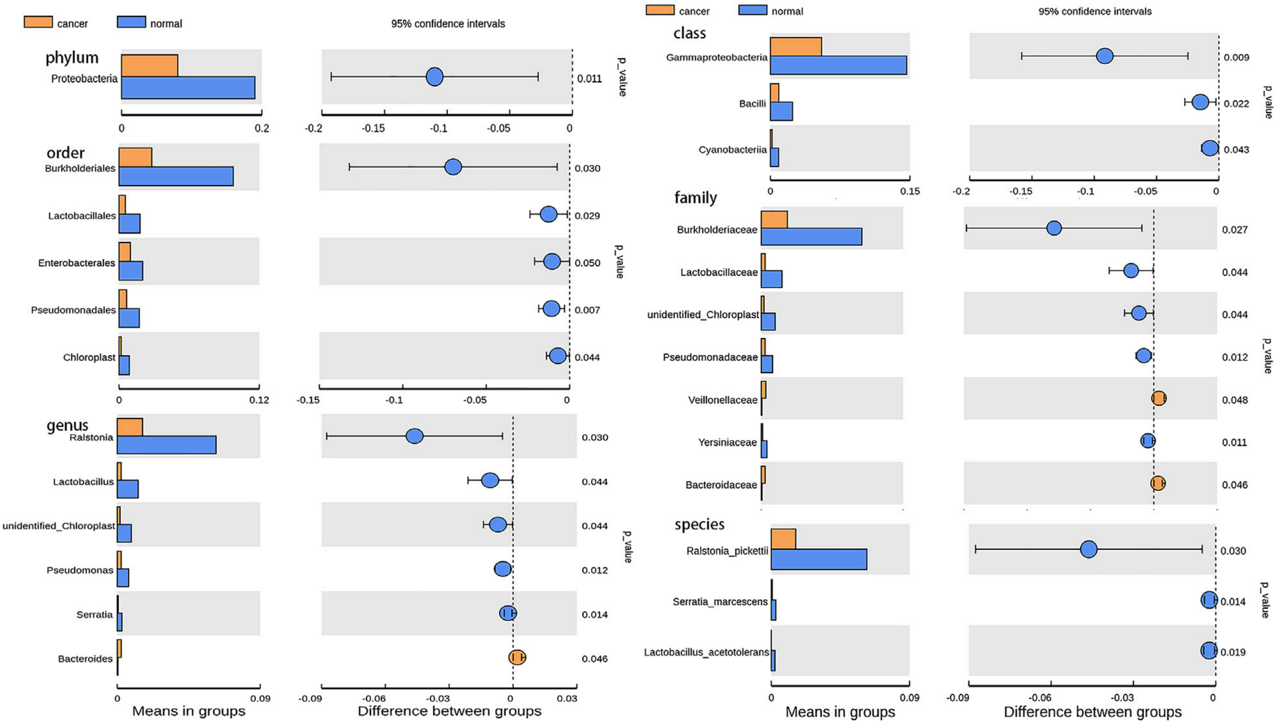


FIGURE 4
Bar plot presents the microbiota with significant differential relative abundance on the phylum, class, order, family, genus, and species levels between lung GGO and paired adjacent normal tissues.

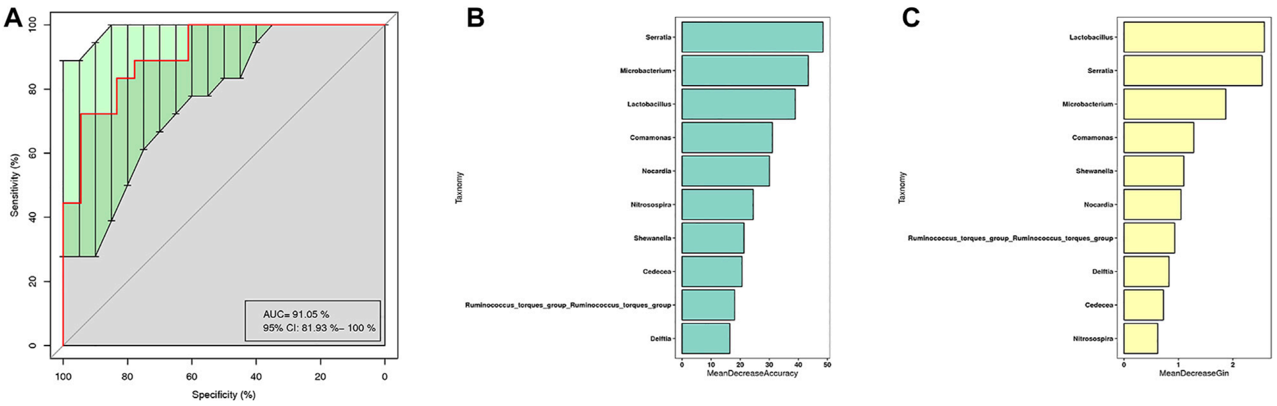


FIGURE 5
Random forest model based on bacterial taxa feature to distinguish GGO tissue and adjacent normal lung tissue. **(A)** Receiver operating characteristic (ROC) curves with the 10 significant differential genera (AUC = 91.05%) to predict GGO and paired adjacent normal tissue. **(B)** Mean decrease accuracy measures the degree of reduction in the accuracy of random forest prediction by changing the value of a variable into a random number. The higher the value, the greater the importance of the variable. **(C)** Mean decrease Gini calculates the influence of each variable on the heterogeneity of observations at each node of the classification tree through Gini index, so as to compare the importance of variables. The higher the value, the greater is the importance of the variable.

very similar to that of the upper respiratory tract, and the oral flora may be the main source of the respiratory tract flora (Dickson and Huffnagle, 2015). In our study, the four main phyla of BALF are *Bacteroidota*, *Proteobacteria*, *Firmicutes*, *Fusobacteriota*, and *Actinobacteria*, which are consistent with the results of other studies on the composition of microbiota in BALF at the phylum level and commonly found in the oral cavity (Jin et al., 2019; Cheng et al., 2020). Cheng et al. (2020) found phylum TM7 and six genera were enriched in the lung cancer group compared with the control group by comparing BALF samples from patients with lung cancer ($n = 32$) and patients with benign lung disease ($n = 22$). However, the same microbiome differences were not found in our study, but we found that there was greater abundance of family *Lachnospiraceae* in BALF of GGO patients. Some studies have found that richness of family *Lachnospiraceae* is related to the low survival rate of lung cancer patients, which seems to indicate that it is also related to early lung cancer such as GGO. Interestingly, *Lachnospiraceae* can produce anti-inflammatory short chain fatty acids (Louis and Flint, 2009), which seems to be inconsistent with the current result. The difference between studies may be caused by differences in the environment, geographical location, and eating habits. In addition, different sampling methods may also be another reason for different results. Furthermore, it is also related to the heterogeneity of each person's lung microbiome.

Our results confirm that BALF is indeed vulnerable to contamination by the upper respiratory tract and oral microbiota. In this study, we also used lung tissue directly obtained from surgery, so that we can not only obtain the actual lung tissue microbiome but also reduce possible oral contamination through sample collection. In this study, our results show that the lung microbiota of cancer patients is different from that of other sites of the body, and the most dominant phylum of lung microbiota is *Proteobacteria*, which is also the main phylum of BALF. Compared with BALF samples, there are some different microbiotas of two kinds of samples. The microbiome of lung tissue samples is more complex and the percentage of main microbiota is lower. However, it is worth noting that the main microbiota of the two samples are similar, also the specific proportion is different, this suggests that the microbiome of lung tissue may also be affected by lower respiratory tract microbiota. Our results are partly consistent with previous studies, which revealed the lung microbiome in lung cancer at the phylum level (Mao et al., 2020). However, compared with previous studies, we did not observe the relative abundance difference of *unclassified Comamonadaceae* and *Propionibacterium* at other taxonomic levels between lung cancer and adjacent tissues (Mao et al., 2020), which indicated that there may be differences in the composition of lung microbiome between GGO and typical lung cancer. However, the microbiome characteristics of GGO are still unclear. A small sample study found that the core microbiotas in GGO tissue are *Mycobacterium*, *Corynebacterium*, and *Negativicoccus* (Ren et al., 2019b). Nevertheless, they did not find the different microbiota between GGO and adjacent normal tissues. Our results were partially the same as a recent study, which explored microbiome diversity through tumor tissues of lung ground-glass nodules and solid nodules (Ma et al., 2021). However this study did not involve with the microbiome of BALF, as well as the relationship of microbiome between BALF and tumor tissues. In our study, we found reduced genera including *Ralstonia*, *Lactobacillus*, *unidentified-Chloroplast*, and *Pseudomonas* in the GGO group, Among them, *Ralstonia pickettii* was found to be a mesothelioma specific microbiota involved in tumor progression (Higuchi et al., 2021), and *Lactobacillus* induce anticancer effect by promoting cancer cell apoptosis and preventing oxidative stress, which is common in probiotics (Badgeley et al., 2021), the effect on GGO can mechanically be interpreted by carcinogenesis due to the decreased genera. Interestingly, the most dominant phylum *Proteobacteria* is also significantly reduced. The results indicate that the microbiota in the local microenvironment may also be involved in the initiation and progression of GGO. In our study, all 10 different bacterial genera were used to distinguish GGO and normal lung tissue through the method of random forest analysis, and the AUC was 91.05%, indicating that these bacterial genera have certain value in discriminating GGO and normal lung tissue.

However, our study also has some limitations. First, the sample size is too small to generate credible evidence. Therefore, larger samples and dynamic longitudinal studies are needed in the future to verify the association between microbiome and different pathological types of lung cancer based on different regions and populations. Second, our studies need to combine bacterial and clinical characteristics to raise the ROC value, which indicates that the combined multidimensional data can better predict lung cancer to a certain extent. Finally, we do not obtain lung tissue samples from healthy patients in this study because it is immoral to obtain lung biopsy from healthy subjects, which is also an unsolvable problem for later researchers.

Conclusion

In conclusion, this is the first time to investigate the microbiome diversity of GGO by BALF combined with lung tissue samples. We found significant differences in the lower respiratory tract and lung microbiome between GGO and the matched non-malignant control group through the BALF and lung tissues. These features may be potential bacterial biomarkers and new targets for GGO diagnosis and treatment.

Data availability statement

The datasets presented in this study can be found in online repositories. The name of the repository and accession number can be found in the following: NCBI; PRJNA843353.

Ethics statement

The studies involving human participants were reviewed and approved by the ethical review board of the First Affiliated Hospital of Zhejiang University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

XL and JH made contributions to experimental conception and design. ZW and JT performed the experiments and wrote the manuscript. LZ, CG, XT, JP, ZZ, and JL provided data analysis and interpretation help. ZW, JT, DM, RZ, and XT contributed to the acquisition and delivery of samples.

Funding

This study was supported by counterpart funding (B20402) of First Affiliated Hospital, College of Medicine, Zhejiang University.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.892613/full#supplementary-material

References

Aberle, D. R., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., Fagerstrom, R. M., et al. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med. Overseas. Ed.* 365 (5), 395–409. doi:10.1056/nejmoa1102873

Badgeley, A., Anwar, H., Modi, K., Murphy, P., and Lakshmikuttyamma, A. (2021). Effect of probiotics and gut microbiota on anti-cancer drugs: Mechanistic perspectives. *Biochimica Biophysica Acta - Rev. Cancer* 1875 (1), 188494. doi:10.1016/j.bbcan.2020.188494

Boursi, B., Mamtani, R., Haynes, K., and Yang, Y. X. (2015). Recurrent antibiotic exposure may promote cancer formation--Another step in understanding the role of the human microbiota? *Eur. J. Cancer* 51 (17), 2655–2664. doi:10.1016/j.ejca.2015.08.015

Budden, K. F., Shukla, S. D., Rehman, S. F., Bowerman, K. L., Keely, S., Hugenholtz, P., et al. (2019). Functional effects of the microbiota in chronic respiratory disease. *Lancet Respir. Med.* 7 (10), 907–920. doi:10.1016/s2213-2600(18)30510-1

Cameron, S. J. S., Lewis, K. E., Huws, S. A., Hegarty, M. J., Lewis, P. D., Pachebat, J. A., et al. (2017). A pilot study using metagenomic sequencing of the sputum microbiome suggests potential bacterial biomarkers for lung cancer. *PLoS One* 12 (5), e0177062. doi:10.1371/journal.pone.0177062

Chang, B., Hwang, J. H., Choi, Y. H., Chung, M. P., Kim, H., Kwon, O. J., et al. (2013). Natural history of pure ground-glass opacity lung nodules detected by low-dose CT scan. *Chest* 143 (1), 172–178. doi:10.1378/chest.11-2501

Cheng, C., Wang, Z., Wang, J., Ding, C., Sun, C., Liu, P., et al. (2020). Characterization of the lung microbiome and exploration of potential bacterial biomarkers for lung cancer. *Transl. Lung Cancer Res.* 9 (3), 693–704. doi:10.21037/tlcr-19-590

Dickson, R. P., Erb-Downward, J. R., Martinez, F. J., and Huffnagle, G. B. (2016). The microbiome and the respiratory tract. *Annu. Rev. Physiol.* 78, 481–504. doi:10.1146/annurev-physiol-021115-105238

Dickson, R. P., and Huffnagle, G. B. (2015). The lung microbiome: New principles for respiratory bacteriology in health and disease. *PLoS Pathog.* 11 (7), e1004923. doi:10.1371/journal.ppat.1004923

Fan, L., Liu, S. Y., Li, Q. C., Yu, H., and Xiao, X. S. (2012). Multidetector CT features of pulmonary focal ground-glass opacity: differences between benign and malignant. *Br. J. Radiol.* 85 (1015), 897–904. doi:10.1259/bjr/33150223

Gomes, M., Teixeira, A. L., Coelho, A., Araújo, A., and Medeiros, R. (2014). The role of inflammation in lung cancer. *Adv. Exp. Med. Biol.* 816, 1–23. doi:10.1007/978-3-0348-0837-8_1

Higuchi, R., Goto, T., Hirotsu, Y., Otake, S., Oyama, T., Amemiya, K., et al. (2021). Streptococcus australis and Ralstonia pickettii as major microbiota in mesotheliomas. *J. Pers. Med.* 11 (4), 297. doi:10.3390/jpm11040297

Hilty, M., Burke, C., Pedro, H., Cardenas, P., Bush, A., Bossley, C., et al. (2010). Disordered microbial communities in asthmatic airways. *PLoS One* 5 (1), e8578. doi:10.1371/journal.pone.0008578

Jin, J., Gan, Y., Liu, H., Wang, Z., Yuan, J., Deng, T., et al. (2019). Diminishing microbiome richness and distinction in the lower respiratory tract of lung cancer patients: A multiple comparative study design with independent validation. *Lung Cancer* 136, 129–135. doi:10.1016/j.lungcan.2019.08.022

Lee, S. H., Sung, J. Y., Yong, D., Chun, J., Kim, S. Y., Song, J. H., et al. (2016). Characterization of microbiome in bronchoalveolar lavage fluid of patients with lung cancer comparing with benign mass like lesions. *Lung Cancer* 102, 89–95. doi:10.1016/j.lungcan.2016.10.016

Liu, H. X., Tao, L. L., Zhang, J., Zhu, Y. G., Zheng, Y., Liu, D., et al. (2018). Difference of lower airway microbiome in bilateral protected specimen brush between lung cancer patients with unilateral lobar masses and control subjects. *Int. J. Cancer* 142 (4), 769–778. doi:10.1002/ijc.31098

Liu, Y., O'Brien, J. L., Ajami, N. J., Scheurer, M. E., Amirian, E. S., Armstrong, G., et al. (2018). Lung tissue microbial profile in lung cancer is distinct from emphysema. *Am. J. Cancer Res.* 8 (9), 1775–1787.

Louis, P., and Flint, H. J. (2009). Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiol. Lett.* 294 (1), 1–8. doi:10.1111/j.1574-6968.2009.01514.x

Ma, Y., Qiu, M., Wang, S., Meng, S., Yang, F., Jiang, G., et al. (2021). Distinct tumor bacterial microbiome in lung adenocarcinomas manifested as radiological subsolid nodules. *Transl. Oncol.* 14 (6), 101050. doi:10.1016/j.tranon.2021.101050

MacMahon, H., Naidich, D. P., Goo, J. M., Lee, K. S., Leung, A. N. C., Mayo, J. R., et al. (2017). Guidelines for management of incidental pulmonary nodules detected on CT images: From the fleischner society 2017. *Radiology* 284 (1), 228–243. doi:10.1148/radiol.2017161659

Maddi, A., Sabharwal, A., Violante, T., Manuballa, S., Genco, R., Patnaik, S., et al. (2019). The microbiome and lung cancer. *J. Thorac. Dis.* 11 (1), 280–291. doi:10.21037/jtd.2018.12.88

Mammen, M. J., and Sethi, S. (2016). COPD and the microbiome. *Respirology* 21 (4), 590–599. doi:10.1111/resp.12732

Mao, Q., Jiang, F., Yin, R., Wang, J., Xia, W., Dong, G., et al. (2018). Interplay between the lung microbiome and lung cancer. *Cancer Lett.* 415, 40–48. doi:10.1016/j.canlet.2017.11.036

Mao, Q., Ma, W., Wang, Z., Liang, Y., Zhang, T., Yang, Y., et al. (2020). Differential flora in the microenvironment of lung tumor and paired adjacent normal tissues. *Carcinogenesis* 41 (8), 1094–1103. doi:10.1093/carcin/bgaa044

Peters, B. A., Hayes, R. B., Goparaju, C., Reid, C., Pass, H. I., Ahn, J., et al. (2019). The microbiome in lung cancer tissue and recurrence-free survival. *Cancer Epidemiol. Biomarkers Prev.* 28 (4), 731–740. doi:10.1158/1055-9965.epi-18-0966

Ramírez-Labrada, A. G., Isla, D., Artal, A., Arias, M., Rezusta, A., Pardo, J., et al. (2020). The influence of lung microbiota on lung carcinogenesis, immunity, and immunotherapy. *Trends Cancer* 6 (2), 86–97. doi:10.1016/j.trecan.2019.12.007

Ren, Y., Huang, S., Dai, C., Xie, D., Zheng, L., Xie, H., et al. (2019). Germline predisposition and copy number alteration in pre-stage lung adenocarcinomas presenting as ground-glass nodules. *Front. Oncol.* 9, 288. doi:10.3389/fonc.2019.00288

Ren, Y., Su, H., She, Y., Dai, C., Xie, D., Narrandes, S., et al. (2019). Whole genome sequencing revealed microbiome in lung adenocarcinomas presented as ground-glass nodules. *Transl. Lung Cancer Res.* 8 (3), 235–246. doi:10.21037/tlcr.2019.06.11

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12, 87. doi:10.1186/s12915-014-0087-z

Tsay, J. J., Wu, B. G., Badri, M. H., Clemente, J. C., Shen, N., Meyn, P., et al. (2018). Airway microbiota is associated with upregulation of the PI3K pathway in lung cancer. *Am. J. Respir. Crit. Care Med.* 198 (9), 1188–1198. doi:10.1164/rccm.201710-2118oc

Yu, G., Gail, M. H., Consonni, D., Carugno, M., Humphrys, M., Pesatori, A. C., et al. (2016). Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features. *Genome Biol.* 17 (1), 163. doi:10.1186/s13059-016-1021-1

Zhang, W., Luo, J., Dong, X., Zhao, S., Hao, Y., Peng, C., et al. (2019). Salivary microbial dysbiosis is associated with systemic inflammatory markers and predicted oral metabolites in non-small cell lung cancer patients. *J. Cancer* 10 (7), 1651–1662. doi:10.7150/jca.28077

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