Identification of unique cytokine expressions in tumor immune microenvironment for various cancer types using single-cell RNA sequencing data via machine learning

*Zhixiang Ren[1,2], Yiming Ren[1], Pengfei Liu[3], Huan Xu[2]**

[1] *Peng Cheng Laboratory, Shenzhen, Guangdong Province, China 518055*

[2] *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism and Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai, China 200237*

[3] *School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong Province, China 528406*

[*] *To whom correspondence should be addressed. Tel: +86 (021) 6425-3739. Email: xuhuan@ecust.edu.cn*

ABSTRACT

Cytokines are small protein molecules exhibit potent immunoregulatory properties, which are known as the essential components of tumor immune microenvironment (TIME). While some cytokines are known to be universally upregulated in TIME, the unique cytokine expression patterns have not been fully resolved in specific types of cancers. In this study, single-cell RNA sequencing (scRNA-seq) data was collected from multiple public repositories, filtered and processed to visualize the cytokine expressions in different immune cells in TIME. The TIME dataset including 39 cancer types was constructed based on the 695 processed immune cell-containing samples. With a random forest model, unique cytokine expression patterns were identified for various cancer categories with high accuracy of nearly 80%. The results will not only boosting the understanding of the cancer type-specific immune modulations in TIME, but also server as an crucial reference for future diagnostic and therapeutic researches in cancer immunity.

Keywords: Tumor immune microenvironment; Cytokines; Machine learning; Single-cell RNA sequencing; Random forest; Immune cells

1. INTRODUCTION

Besides malignant cells, there are many other components in tumors, including stroma cells, immune cells, hormones, cytokines and many other biologically-active molecules, forming a well-organized complex system [1]. Tumor immune microenvironment (TIME) consists of the immunological components within tumors, which is known to be associated with tumor development, progression, metastasis and immune escape [2]. Immune cells in TIME, including monocytes, macrophages, neutrophils, dentritic cells (DCs), natural killer (NK) cells, lymphocytes, and rare progenitors like common myeloid progenitors (CMPs) or hematopoietic stem cells (HSCs), interact with each other and tumor cells through cell signaling molecules, constituting a complicated regulatory network in TIME [3, 4]. Identifying the expression patterns of these signaling molecules not only reveals immune cell regulations in TIME, but also provides essential information for the modeling of TIME immune network.

Cytokines are small but essential proteins that exert potent immunoregulatory roles in physiological and pathological conditions. More than 200 cytokines have been identified so far, which can be categorized into different families, including interleukins (ILs), colony stimulating factor (CSFs), interferons (IFNs), tumor necrosis factors (TNFs), growth factor (GFs) and chemokines [5]. Considered as the hormones of the immune system, many cytokines, such as TNF-α, IL-1β, IL-2, IL-6, IL-12 and IL-15, have already been utilized in cancer immunotherapies either to enhance growth inhibitory effects or suppress tumor-promoting actions [6, 7]. While a histology-independent uniform cytokine cascade was noticed in different types of cancers, unique cytokine expression patterns are also being identified in specific cancer types [8, 9], which may be developed as potential diagnostic biomarkers or therapeutic targets [6, 10, 11].

Recent advances in single-cell RNA sequencing (scRNA-seq) technique have enabled us to study gene expressions at RNA level in individual cells isolated from tumor tissues [12]. The scRNA-seq data deposited in various public repositories can be utilized for comprehensive analysis of TIME. For example, CancerSCEM launched an online analytical platform of TIME based on the scRNA-seq data of 208 cancer samples from 28 projects [13]. TMExplorer created a R package for gene expression analysis in the 48 processed datasets representing 26 different human cancer types and 4 different mouse cancer types [14]. Although these studies provided useful tools for visualization of TIME data, the outputs were too general to reflect the specific cytokine expressions in individual cancer types. In order to identify the precise cytokine expression patterns in TIME of different cancers, scRNA-seq data was collected from various public repositories and processed with carefully designed workflow (Figure 1). Cancer type-unique cytokine expressions were identified via machine learning (ML) after the visualization of cancer type-specific immune cell-cytokine relations. So far, this is the very first study focusing on unique cytokine expression patterns in TIME of various types of cancers.

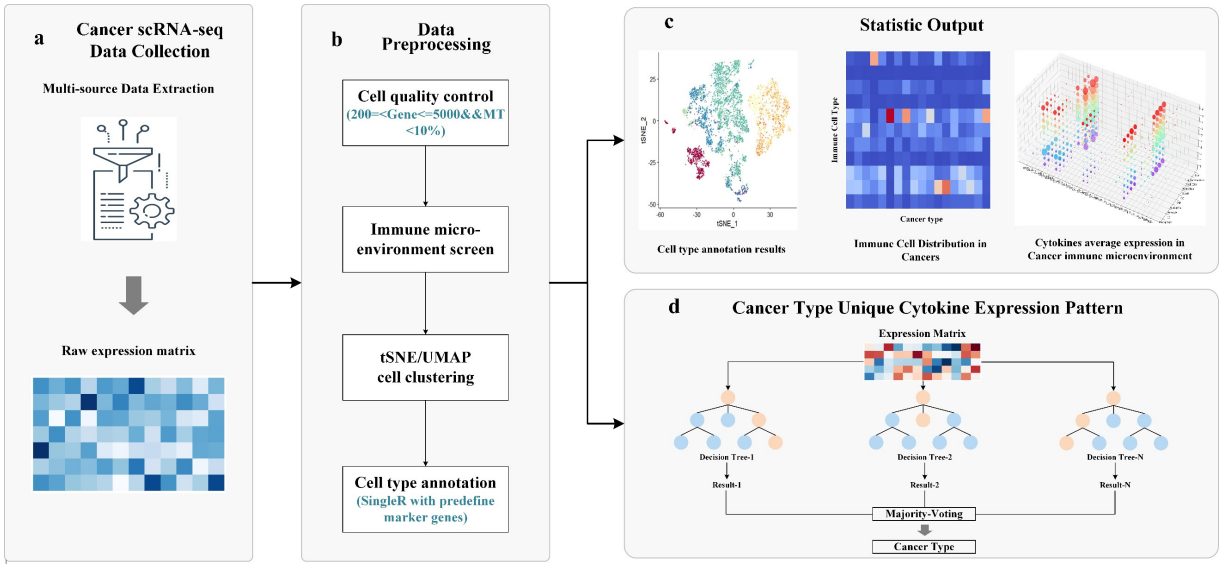

Figure 1. The workflow of the current work, including scRNA-seq data collection, data processing, statistic output and recognition of cancer type-unique cytokine expression patterns.

2. METHODS

2.1. Cancer scRNA-seq data collection and processing

The dataset used in this study was collected from multiple public scRNA-seq databases, including GEO (https://www.ncbi.nlm.nih.gov/geo/) [15], ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress) [16], Single Cell Expression Atlas (https://www.ebi.ac.uk/gxa/sc/home) [17], and ZENODO (https://zenodo.org/). A total 695 human tumor tissue samples were collected and processed from 117 scRNA-seq projects covering 39 human cancers (Table S1). The expression matrix of raw count and R package Seurat v4.3.0 (https://satijalab.org/seurat/) were utilized for single-cell analysis workflow, including cell quality control (200 ≤ nfeatures ≤ 5000 and MT < 10%), PCA (principal component analysis) dimension reduction, UMAP (uniform manifold

approximation and projection) and tSNE (t-distributed stochastic neighbor embedding) clustering with specific parameters and resolutions [18]. After removing low-quality cells, the TIME was initially screened based on marker genes. According to the marker genes on CancerSCEM documentation [13], 7 immune cell types, including B cell, GMP, HSC, Macro/Mono/DC, Neutrophil, NK cell, T cell, were defined to comprise the TIME of this study. The marker genes utilized for immune cell identification in TIME are shown in Table S2. The processed datasets can be accessed on ZENODO (https://sandbox.zenodo.org/record/1180710)

2.2. Immune cell annotation in TIME

Currently, other than manual annotation, there is no standard method for immune cell identification in single-cell analysis. However, manual operation based on a couple of literature is susceptible to subjectivity. SingleR (single cell recognition) is a novel computational framework for cell annotation of scRNA-seq data by reference to aggregate transcriptomes, which has become one of the most trusted tools for cell type recognition in recent years [19]. To determine the diverse immune cell types in TIME, the cell type annotation was implemented with SingleR v2.0.0 (https://github.com/dviraran/SingleR). A total of 11 types of human immune cells were focused in this study, which were B cell, CMP, HSC, DC, Monocyte, Macrophage, Neutrophil, NK cell, T cell: CD4+, T cell: CD8+ and T cell: gamma-delta (Figure S1). Human Primary Cell Atlas (HPCA) [20] consists of 713 microarray samples derived from 37 primary cell types and 157 subtypes of human cells, including all the cell types of interest in TIME. Therefore, HPCA was adopted as the reference dataset to assist in cell type recognition, and only 368 samples contained the 11 cell types of interest were retained. In addition, to improve the accuracy of immune cell recognition, a set of typical marker genes for the concerned cell types in TIME were

customized according to previous studies (Table 1) [19, 21, 22]. The minimum difference median (> 0.1), which quantifies the difference between the assigned label and all other labels, was used to filter the unreliable annotation. Figure S1 represents the recognition effect of SingleR, and the average error rate is less than 5% across all samples.

Table 1. Marker genes for cell type recognition. Customize a set of marker genes for spearman correlation to replace the differential expression genes between samples as determined by SingleR.

Cell Type	Marker Genes
B cell	CD19, MS4A1, BANK1, BLK, IRF8, ABCB4, ABCB9, AFF4, AIDA, AIM2
CMP	CD38, KIT, ADK, CD123, ALDH4A1, ANXA1, AP3S1, APLP2, APPL1, AREG, ASPM, CDKN3, CLSPN, DEPDC7, MCM10, MUCB2, SDC4, RMI2
Neutrophils	ADGRG3, CXCL8, FCGR3B, MNDA, USP10, CSF3R, ANXA3, AQP9, BTNL8, LGALS13, G0S2, NFE4, IL5RA
NK cell	FCGR3A, KLRB1, KLRD1, NKG7, XCL1, XCL2, NCR3, NCR1, CD247, GZMB, KLRC1, KLRK1
T cell:CD4+	CD40LG, CD4, CD3E
T cell:CD8+	GZMK, CD8A, CD8B
T cell:gamma-delta	CD3E, TRDC
Monocyte	FCN1, CD14, VCAN
Macrophage	CSF1R, FCGR3, ADGRE1, MS4A7
DC	NAPSA, TRAF1, CD209A, FLT3
HSC	CD34, ITGA5, PROM1, CD105, VCAM1, CD164, THY1, KIT, ACE, CMAH, ABCG2, CD41, ALDH1A1, BMI1

2.3. Identification of cancer type-unique cytokine expressions in TIME with ML

To investigate the specific cytokine expression patterns in various types of cancers, we employed the constructed dataset in conjunction with ML. Random forest was applied for cytokine identification with the Python package Scikit-learn [23], and F1 score was introduced to evaluate the classification accuracy

of each category. Since variations in the quantity of cells in each processed scRNA-seq sample may result in classification bias, the expression matrix of each sample was converted into a 106×11 matrix by averaging the expression of 106 cytokines across the 11 immune cell types, which ensured the uniformity of characteristics and eliminated the cell number differences between samples. In addition, the proportion of immune cells in each sample was appended to the final dimension of the training data, as the ratio of immune cells was regarded as an important feature for cancer classification (Figure 2). Given the substantial variance in the amount of samples across 39 cancer categories, the 10 cancers with the largest sample quantity were selected to improve the generalizability of the ML model.

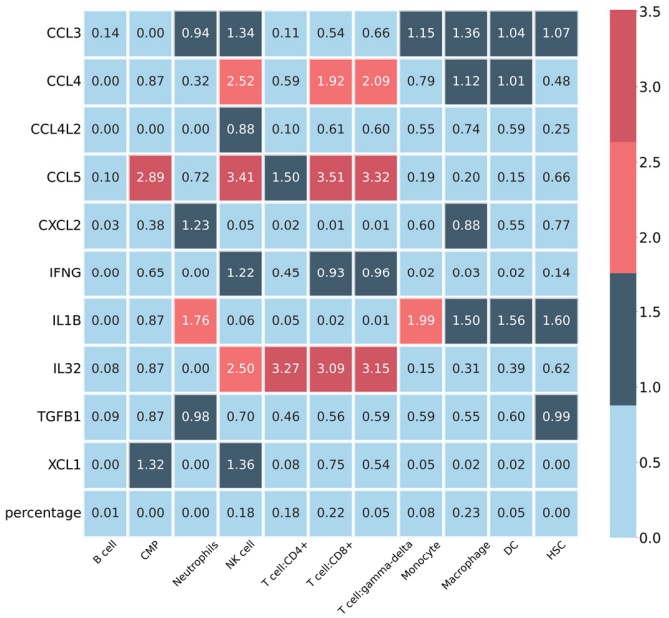

Figure 2. A demo of training data represents the average expression of selected cytokines in specific immune cell types. The rows contain two types of information: the selected cytokines and the proportion of each type of immune cell in the sample. The columns represent the 11 types of immune cells involved in the study.

3. RESULTS

3.1. Data statistics and visualization of TIME

After the cell type recognition with SingleR, tSNE and UMAP were applied to visualize the distribution of cells in the TIME (Figure 3A and B).!Jn order to investigate cytokine expressions in TIME, highly expressed cytokines in various types of immune cells (percentage>10%) in each sample were listed in a bubble chart (Figure 3C). A total of 18 cancer categories were chosen for further analysis based on their sample sizes (amount of samples⩾10). The immune cell types in the TIME of the selected cancer samples and the top 20 expressed cytokines in the 18 types of cancers were also demonstrated in Figure S2. In addition, the correlations between cytokine expressions and immune cell types in the 10 cancer categories with the biggest sample sizes were summarized in a three-dimensional graph, visualizing the general cytokine expression patterns in the TIME of various cancer types (Figure 4). The average expressions of cytokines across all cancer types were also analyzed, identifying the commonly expressed cytokines in TIME, such as CCL4, IL32, CCL5, CCL3, CCL4L2, IL1B, TGFB1, CCL3L1, CXCL2, and TNFA, which were being ruled out for the recognition of unique cytokines for certain cancer types (Figure 5).

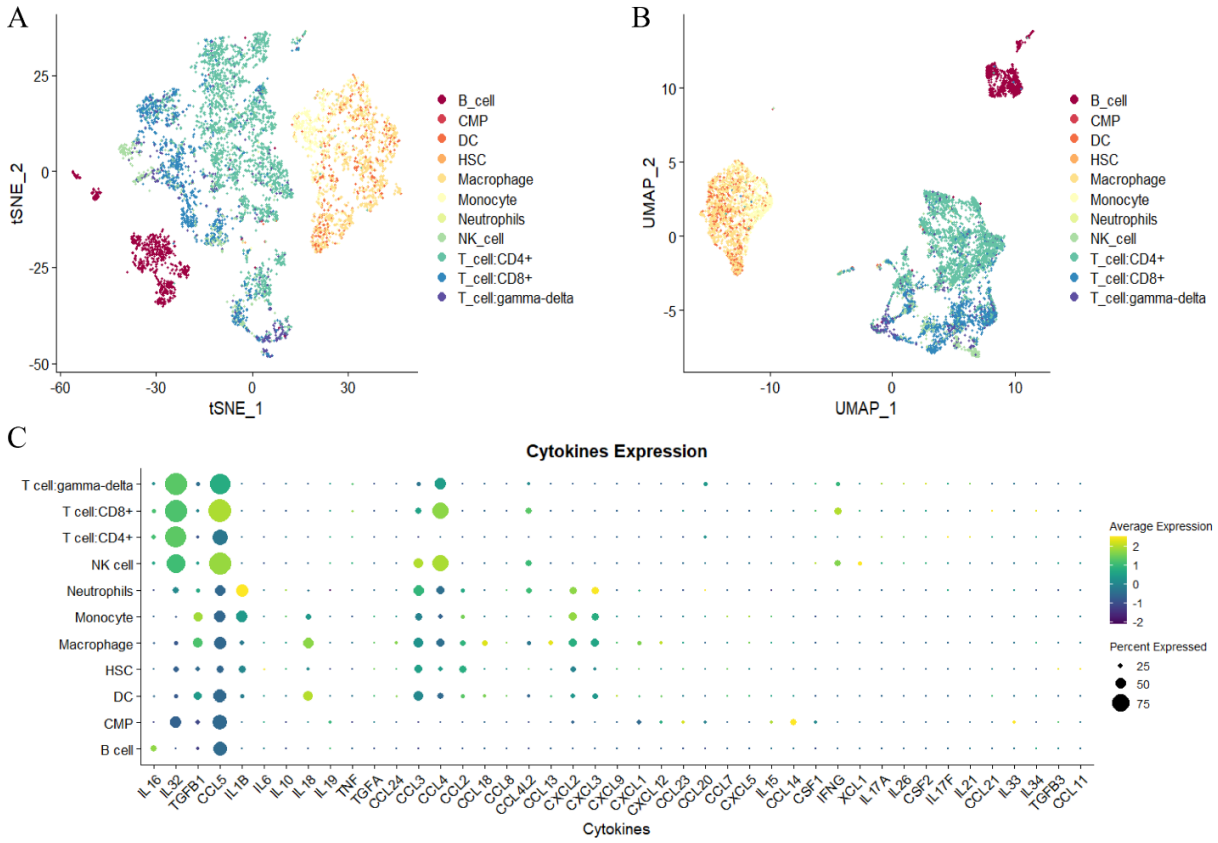

Figure 3. A and B, the tSNE (A) and UMAP (B) clustering results represent the distribution of different types of immune cells in the microenvironment. C, the average expression of cytokines in each immune cell type. Cytokines with the expression percentage lower than 10% were filtered out.

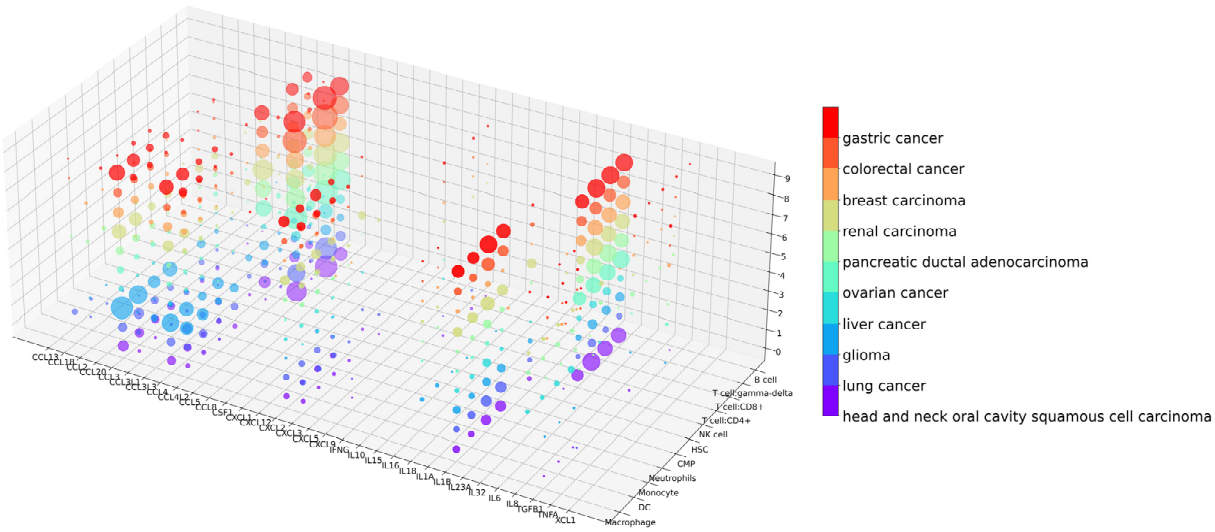

Figure 4. Three-dimensional graph representing the expression of the top 20 cytokines in immune cells of the 10 cancers with the most samples.

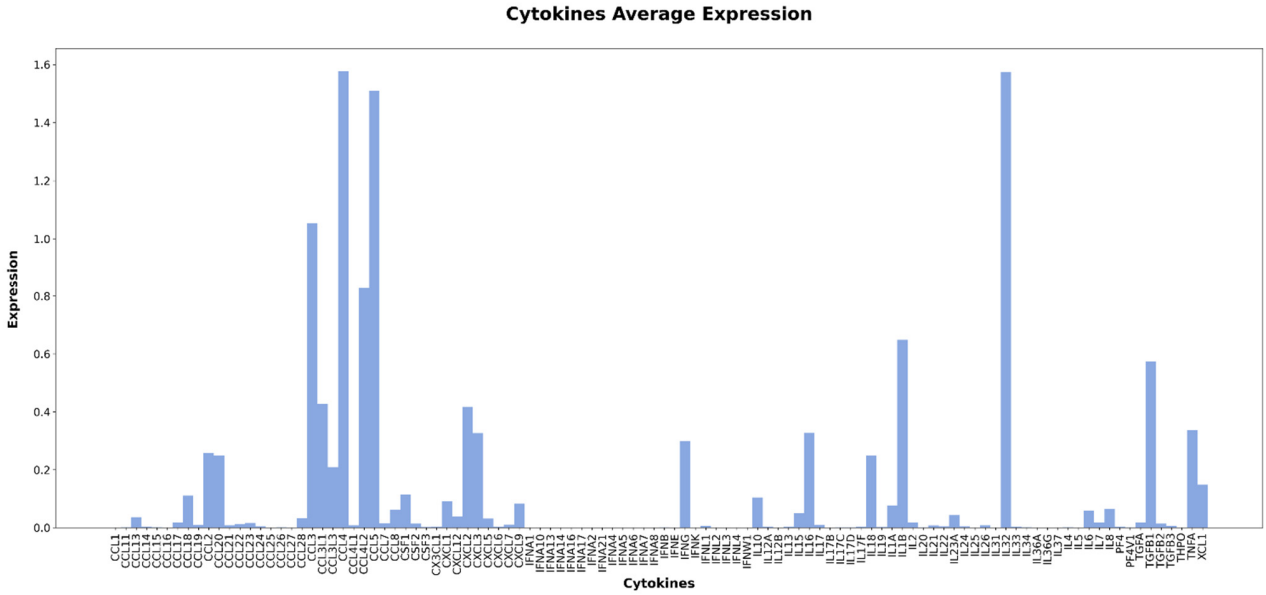

Figure 5. The distribution of cytokines average expression across all samples, with the top-10 highly-expresed cytokines being CCL4, IL32, CCL5, CCL3, CCL4L2, IL1B, TGFB1, CCL3L1, CXCL2, and TNFA.

3.2. Identification of specific cytokines with ML

After the performance evaluation of decision tree, random forest, and multi-layer perceptron (MLP) models (Figure S3), random forest was applied in this study to verify the identification efficiency of the model in 3 to 10 types of cancers. The results of random forest classifying with 106 cytokines demonstrated that the accuracy of the top three cancer types with the most samples could achieve a high precision of 92% (Table 2; Figure S4). Although the accuracy declined progressively as the number of cancer samples decreased, the classification results of 10 cancer types still achieved a remarkable performance of nearly 80%. Moreover, to explore highly-distinct cytokine expressions associated with each cancer type, the specific cytokines were extracted using the FindMarkers function of the scran v1.26.1 (https://bioconductor.org/packages/release/bioc/html/scran.html). The generated matrix was described as the concatenation of multiple samples corresponding to each cancer type. By continuously adjusting the threshold, we were able to identify multiple groups of specifically-expressed cytokines. In the threshold selection part, only the top 10 differential genes with a threshold (p-value < 0.01) were considered due to the high differential expression. To ensure that at least one specific cytokine was identified for each type of cancer, the minimum number of specific genes was set at 41. In addition, 65 was set as the upper limit, since the model effect diminished when the number of specific genes exceeded 65. The identified unique cytokines in each cancer type with 41 or 65 differential genes were listed in Table 3 and 4. The application with 41 cytokines resulted in a reduction in classification accuracy but still had a minimum of 71.5%. However, as the number of selected cytokines increased, the model's accuracy significantly dropped, indicating that the number of cytokines should be limited to an appropriate range in this model (Figure S4). In addition, the accuracy of ovarian cancer was consistently low, and the F1-score decreased considerably from 0.69 to 0.57 when specific cytokines were used, suggesting the complexity

of cytokine expressions in the particular cancer type. For other cancer types, the identified cytokines with lowest p-values were consistent with threshold alterations, which were summarized in Figure 6. These results will not only benefit for the understanding of immune cell regulations in TIME, but also provide valuable information for cancer immunity-related diagnostic and therapeutic researches in the future.

Table 2. Classification accuracy of random forest with different cytokines. The table shows the F1 score of cancer classification based on different cytokines. According to the amount of samples, the top 10 cancer types with the most samples were chosen, which are breast carcinoma, head and neck oral cavity squamous cell carcinoma, lung cancer, gastric cancer, colorectal cancer, ovarian cancer, renal carcinoma, glioma, pancreatic ductal adenocarcinoma, liver cancer. F1@106 represents the F1 score of classification with the whole 106 cytokines.

Number of categories	F1@106	F1@41	F1@61	F1@65
3	92.63%	90.52%	90.95%	91.81%
4	89.14%	86.06%	88.14%	87.45%
5	86.31%	81.94%	84.88%	85.2%
6	83.23%	77.34%	78.13%	79.95%
7	80.66%	74.06%	78.63%	79.33%
8	78.81%	74.17%	75.9%	77%
9	77.61%	73.45%	74.88%	76.31%
10	78.01%	71.55%	75.62%	75.81%

Table 3. 41 differential cytokines of top 10 cancer types obtained based on scran's FindMarkers function with relatively strict threshold. Only top 10 genes with p-value < 0.01 were selected.

Cancer Type	Differential Cytokines
breast carcinoma	IL21, CCL19, CCL3L3, CXCL9, CCL8, IFNL4, IL17B CCL22, CCL21, IL17F

head and neck oral cavity squamous cell carcinoma	CCL18, CCL13, CCL23, CXCL5, IL24, CCL17, IL21, IL37, CCL8
lung cancer	CCL24, IL17F, IL17D
gastric cancer	IL17, IL22, IL17F, CCL15, IL26
colorectal cancer	XCL1
ovarian cancer	CCL14, CX3CL1
renal carcinoma	CCL3L1, CCL2, IL1A, IL18, IL1B, IL10, IL6, CCL8, CCL4L2, CCL20
glioma	IL8, CCL18, CCL4L1
pancreatic ductal adenocarcinoma	PF4, CXCL7, CCL16, PF4V1, CCL15, THPO, IL8
liver cancer	

Table 4. 65 differential cytokines of top 10 cancer types obtained based on scran's FindMarkers function with relatively strict threshold. Only top 10 genes with p-value < 0.01 were selected.

Cancer Type	Differential Cytokines
breast carcinoma	CCL3L3, CXCL9, CCL19, IL21, IL10, CCL8, IL16, IL7, CCL21, CCL2
head and neck oral cavity squamous cell carcinoma	CCL22, CCL21, IL26, CCL17, CCL24, CSF2, IL17F, CCL19, IL17, IFNB
lung cancer	CCL18, CCL13, CCL23, CXCL5, CCL17, CCL8, IL24, CCL2, IL21, CXCL1
gastric cancer	CCL24, IL24, IL17F, IL17D, IL4, IL22, IL17, IL23A, IFNL2, IFNA5
colorectal cancer	IL17, IL22, IL26, CCL15, IL17F, CSF2, CCL11
ovarian cancer	XCL1, CCL5, CCL4, CSF1, IFNG, IL16, IL32, IFNL1, IL5, CCL11
renal carcinoma	IL10, CCL14, CX3CL1, CCL3, CXCL12, IL4, CCL4L1, IL12A, IL17B, THPO
glioma	CCL3L1, CCL2, IL18, IL1B, IL1A, CCL4L2, CCL3, IL6, CCL20, IL10
pancreatic ductal adenocarcinoma	IL8, CCL18, CCL4L1, CCL13, CXCL5, CCL23, TGFA, CXCL6, CCL28, TGFB2
liver cancer	CXCL7, PF4, CCL16, PF4V1, IL8, CCL15, THPO, CCL19, CCL21, CCL25

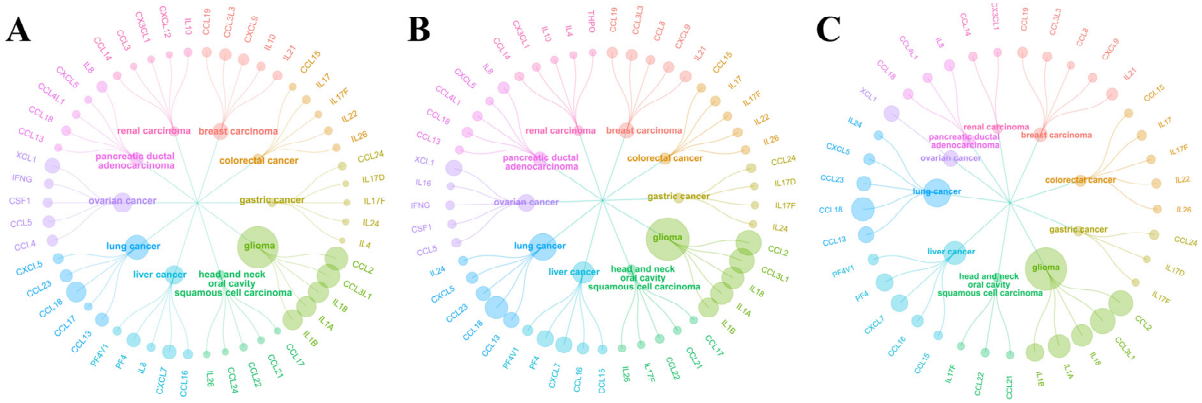

Figure 6. A, B, C show the top 5 immune regulatory cytokines identified in the 10 cancers with the highest sample counts under three different threshold values

4. DISCUSSION

Single-cell technology has revolutionized the understanding of the heterogeneity and functional diversity of individual cells, bringing tremendous opportunities to biology and precision medicine. Constructing the single-cell TIME dataset is of great significance for studying applications, such as tumor heterogeneity, immune system, and drug screening. In this study, we devised a custom Python script to effectively gather scRNA-seq data pertaining to the TIME from a diverse range of cancer patients, primarily drawing from public databases, such as GEO. Our method, which integrated Python libraries like BeautifulSoup and requests with parallel techniques using the concurrent futures library, facilitated simultaneous data collection from multiple webpages, significantly expediting the data collection process and resulting in a comprehensive dataset encompassing over 30 cancer types. Conversely, CancerSCEM [13] and TISCH [24] databases acquire their scRNA-seq data from publicly accessible resources, including GEO, GSA and ArrayExpress, as well as published studies. Although these databases offer

valuable resources, the quantity of datasets they incorporated was comparatively limited. Our approach presented more efficient means of data acquisition, enabling the creation of a refined dataset focused exclusively on tumor samples with accompanying information about tissue origin. In terms of data processing, we established an end-to-end pipeline oriented towards TIME, including cell quality control, immune microenvironment screening, clustering, and cell-type annotation. At the cell-type annotation stage, recognition of immune cells was mainly focused in this study. Unlike the complex mode of combined annotation tools and manual operations, we only used SingleR's recognition algorithm and regard marker genes as hyperparameters. After multiple finetunes, a set of highly differential marker genes was determined which achieved good recognition results on the dataset. The specificity of the data collection and processing strategy empowered us to generate accurate insights into the TIME and its involvement across various cancer types, ultimately contributing to the continued advancement of cancer immunity research.

In recent years, artificial intelligence (AI) techniques such as ML and deep learning have redefined the methods and capabilities for analyzing large-scale data, including scRNA-seq results. Many studies have demonstrated the feasibility of using AI techniques for scRNA-seq data analysis. For example, SAUCIE applied autoencoder to a dataset of 11 million T cells from 180 dengue fever patients, identifying clustering features of acute dengue infection and stratified immune response to dengue fever [25]. Similarly, scGNN combined graph neural network (GNN) with scRNA-seq data to identify 10 neuronal clusters and cell-type-specific markers in Alzheimer's disease [26]. SpaGCN, based on spatially resolved transcriptomics data, identified cancerous and non-cancerous regions of human primary pancreatic tumors and identified two marker genes that distinguish cancerous regions [27]. Dohmen et al.

used ML models to identify tumor cells and normal cells in tumor single-cell data [28]. The present study focused on studying the TIME by combining ML methods with scRNA-seq data. A TIME dataset containing 39 cancers was constructed and random forest was trained to model the relationship between immune regulatory cytokines and cancer types, achieving high accuracy of nearly 80% in the 10 cancers with the highest amount of sample (Figure S4). The results clearly demonstrated that regardless of the universally-expressed cytokines in TIME of all cancers, unique cytokine expression patterns could be recognized for each particular cancer type, which might be further developed as treatment targets or diagnostic markers. In addition, the results provided critical information for future explorations on the immune cells and signaling pathways related to these cytokines. Although we tried to gather as many scRNA-seq samples as possible, the study was still limited by data scale and was difficult to achieve high performance in more complicated TIME visualization tasks. Follow-up studies will continue to extend the dataset and apply advanced deep learning models to model the TIME regulations for cancer diagnosis and treatment.

ACKNOWLEDGEMENT

We appreciate Yue Zhou from Peng Cheng Laboratory for the technical advice and Zhaoqian Ding, Ying Xie, Tingqian Wang from East China University of Science and Technology for the collection of TCM prescriptions from books.

FUNDING INFORMATION

This work was funded by National Natural Science Foundation of China [grant No. 42177417]. The project was supported by the Peng Cheng Laboratory and Peng Cheng Cloud-Brain. This work was also funded by Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism (Shanghai Municipal Education Commission).

DISCLOSURE OF INTEREST

The authors report no conflict of interest.

REFERENCES

1. Gajewski TF, Schreiber H, Fu YX. Innate and adaptive immune cells in the tumor microenvironment. Nat Immunol. 2013 Oct;14(10):1014-22.
2. Fu T, Dai LJ, Wu SY, Xiao Y, Ma D, Jiang YZ, Shao ZM. Spatial architecture of the immune microenvironment orchestrates tumor immunity and therapeutic response. J Hematol Oncol. 2021 Jun 25;14(1):98.
3. Marzagalli M, Ebelt ND, Manuel ER. Unraveling the crosstalk between melanoma and immune cells in the tumor microenvironment. Semin Cancer Biol. 2019 Dec;59:236-250.
4. Li L, Yu R, Cai T, Chen Z, Lan M, Zou T, Wang B, Wang Q, Zhao Y, Cai Y. Effects of immune cells and cytokines on inflammation and immunosuppression in the tumor microenvironment. Int Immunopharmacol. 2020 Nov;88:106939.
5. Dinarello CA. Historical insights into cytokines. Eur J Immunol. 2007 Nov;37 Suppl 1(Suppl 1):S34-45.

6. Propper DJ, Balkwill FR. Harnessing cytokines and chemokines for cancer therapy. Nat Rev Clin Oncol. 2022 Apr;19(4):237-253.

7. Berraondo P, Sanmamed MF, Ochoa MC, Etxeberria I, Aznar MA, Pérez-Gracia JL, Rodríguez-Ruiz ME, Ponz-Sarvise M, Castañón E, Melero I. Cytokines in clinical cancer immunotherapy. Br J Cancer. 2019 Jan;120(1):6-15.

8. Lippitz BE. Cytokine patterns in patients with cancer: a systematic review. Lancet Oncol. 2013 May;14(6):e218-28.

9. Autenshlyus AI, Studenikina AA, Varaksin NA, Lyakhovich VV. Cytokine Production by Tumor Bioptate at Different Pathological Prognostic Stages in Breast Cancer. Dokl Biochem Biophys. 2021 Mar;497(1):86-89.

10. Bellesoeur A, Torossian N, Amigorena S, Romano E. Advances in theranostic biomarkers for tumor immunotherapy. Curr Opin Chem Biol. 2020 Jun;56:79-90.

11. Alfaro C, Sanmamed MF, Rodríguez-Ruiz ME, Teijeira Á, Oñate C, González Á, Ponz M, Schalper KA, Pérez-Gracia JL, Melero I. Interleukin-8 in cancer pathogenesis, treatment and follow-up. Cancer Treat Rev. 2017 Nov;60:24-31.

12. Slovin S, Carissimo A, Panariello F, Grimaldi A, Bouché V, Gambardella G, Cacchiarelli D. Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. Methods Mol Biol. 2021;2284:343-365.

13. Zeng J, Zhang Y, Shang Y, Mai J, Shi S, Lu M, Bu C, Zhang Z, Zhang Z, Li Y, Du Z, Xiao J. CancerSCEM: a database of single-cell expression map across various human cancers. Nucleic Acids Res. 2022 Jan 7;50(D1):D1147-D1155.

14. Christensen E, Naidas A, Chen D, Husic M, Shooshtari P. TMExplorer: A tumour microenvironment single-cell RNAseq database and search tool. PLoS One. 2022 Sep 9;17(9):e0272302.

15. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.

16. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, Sarkans U, Brazma A. ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res. 2019 Jan 8;47(D1):D711-D715.

17. Papatheodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S, Fonseca NA, Füllgrabe A, Green M, Huang N, Huerta L, Iqbal H, Jianu M, Mohammed S, Zhao L, Jarnuczak AF, Jupp S, Marioni J, Meyer K, Petryszak R, Prada Medina CA, Talavera-López C, Teichmann S, Vizcaino JA, Brazma A. Expression Atlas update: from tissues to single cells. Nucleic Acids Res. 2020 Jan 8;48(D1):D77-D83.

18. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29.

19. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019 Feb;20(2):163-172.

20. Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics. 2013 Sep 20;14:632.

21. Pizzolato G, Kaminski H, Tosolini M, Franchini DM, Pont F, Martins F, Valle C, Labourdette D, Cadot S, Quillet-Mary A, Poupot M, Laurent C, Ysebaert L, Meraviglia S, Dieli F, Merville P, Milpied P, Déchanet-Merville J, Fournié JJ. Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRVδ1 and TCRVδ2 γδ T lymphocytes. Proc Natl Acad Sci U S A. 2019 Jun 11;116(24):11906-11915.

22. Li CM, Shapiro H, Tsiobikas C, Selfors LM, Chen H, Rosenbluth J, Moore K, Gupta KP, Gray GK, Oren Y, Steinbaugh MJ, Guerriero JL, Pinello L, Regev A, Brugge JS. Aging-Associated Alterations in Mammary Epithelia and Stroma Revealed by Single-Cell RNA Sequencing. Cell Rep. 2020 Dec 29;33(13):108566.

23. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G. Machine learning for neuroimaging with scikit-learn. Front Neuroinform. 2014 Feb 21;8:14.

24. Sun D, Wang J, Han Y, Dong X, Ge J, Zheng R, Shi X, Wang B, Li Z, Ren P, Sun L, Yan Y, Zhang P, Zhang F, Li T, Wang C. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. Nucleic Acids Res. 2021 Jan 8;49(D1):D1420-D1430.

25. Amodio M, van Dijk D, Srinivasan K, Chen WS, Mohsen H, Moon KR, Campbell A, Zhao Y, Wang X, Venkataswamy M, Desai A, Ravi V, Kumar P, Montgomery R, Wolf G, Krishnaswamy S.

Exploring single-cell data with deep multitasking neural networks. Nat Methods. 2019 Nov;16(11):1139-1145.

26. Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, Wang C, Fu H, Ma Q, Xu D. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat Commun. 2021 Mar 25;12(1):1882.
27. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, Lee EB, Shinohara RT, Li M. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods. 2021 Nov;18(11):1342-1351.
28. Dohmen J, Baranovskii A, Ronen J, Uyar B, Franke V, Akalin A. Identifying tumor cells at the single-cell level using machine learning. Genome Biol. 2022 May 30;23(1):123.