ISSN 2706-6819 Vol.5, Issue 10: 81-87, DOI: 10.25236/IJFM.2023.051013

Dissecting heterogeneity and immune cell populations in non-small cell lung cancer by single-cell RNA sequencing

Tao Yu[1,a,#], Xuehan Gao[2,b,#], Jueyi Zhou[3,c], Liping Zhao[3,d], Jihong Feng[3,e,*]

[1]Department of Respiratory and Critical Care Medicine, People's Hospital of Inner Mongolia Autonomous Region, Hohhot, China
[2]Department of Immunology, Zunyi Medical University, Zunyi, China
[3]Department of Oncology, The Sixth Affiliated Hospital of Wenzhou Medical University, Lishui People's Hospital, Lishui, China
[a]yutao1972858@sina.com, [b]gxh980312@163.com, [c]1193489192@qq.com, [d]672648511@qq.com, [e]jh_f@163.com
([#]co-first author) These authors contributed equally to this work
[*]Corresponding author

Abstract: *Lung cancer is the most common and aggressive cancer and the leading cause of cancer-related death worldwide, with non-small cell lung cancer (NSCLC) being the most common type. However, the issue of tumor heterogeneity in non-small cell lung cancer has received increasing attention and is not currently addressed at single-cell resolution. In this study, integrated single-cell RNA sequencing (scRNA-seq) samples from Non-Small-Cell Lung Cancer (NSCLC) samples and paracancerous control samples were downloaded from the high-throughput Gene Expression Omnibus (GEO) data and batch RNA-seq data for analysis. Three NSCLC cell subsets in different differentiation states were compared and analyzed. GSEA-GO analysis predicts the biological functions and pathways of differentiation-related genes. The sequencing results of a total of 4320 cells from 11 NSCLC samples and 5 paracancerous lung tissue samples were obtained from the GEO database. After data standardization and data filtering, all cells were subjected to unsupervised clustering to obtain 3 different clusters, which were visualized after dimensionality reduction through T-SNE, and 10 differential marker genes were analyzed and screened, which can be clustered in different clusters. Gene set enrichment analysis found that CDRG was significantly associated with immune regulation and immune response, and 278 NSCLC cell differentiation-related genes (CDRG) were identified. Our study identified NSCLC cells with distinct differentiation characteristics based on single-cell sequencing data from GEO, emphasizing the important role of cell differentiation in predicting the clinical outcome of NSCLC patients and their potential response to immunotherapy.*

Keywords: *Non-small cell lung cancer, tumor heterogeneity, immune prognosis, single-cell sequencing, GEO database, GSEA analysis*

1. Introduction

Non-small cell lung cancer (NSCLC) is the main cause of lung cancer-related death. About 85% of new lung cancer patients are pathologically classified as NSCLC, and the 5-year survival rate of this pathological type is less than 16% [1]. NSCLC not only has unique pathological characteristics at the tissue level but also exhibits obvious tumor heterogeneity at the cellular and molecular levels [2]. During the occurrence and development of NSCLC, the evolution of tumor cells gradually diversifies. The corresponding tumor microenvironment is infiltrated by a variety of immune-related components, which leads to more immune heterogeneity in NSCLC. This heterogeneity varies with tumor progression and immunotherapy intervention. And there are changes in different dimensions [3, 4]. Therefore, an in-depth study of tumor immune heterogeneity is crucial for the development of effective treatment of NSCLC [5]. However, lung cancer treatment entered the immune era as early as 5 years ago [6]. However, biomarkers and predictors that affect the prognosis of patients with immunotherapy have not been fully elucidated, and existing predictive models are far from satisfactory.

The rapid development of related detection technologies such as single-cell multi-omics sequencing has prompted RNA-based detection to gradually become a powerful method for studying tumor immune

ISSN 2706-6819 Vol.5, Issue 10: 81-87, DOI: 10.25236/IJFM.2023.051013

heterogeneity. This technology mainly describes cell state transformation and functional clustering by comprehensively studying the characteristics of tumor sample genomes [7]. Single-cell sequencing technology can effectively predict the differentiation trajectories of tumor-related cells. Based on the predicted cell differentiation status, researchers further study tumor cell subsets, and in-depth reveal cell-related tumorigenic mechanisms and pathways [8]. According to current reports, single-cell scRNA-seq-based methods have accelerated the study of immune heterogeneity in various tumors, such as acute myeloid leukemia, breast cancer, pancreatic ductal adenocarcinoma, and malignant melanoma [9-11].

In this study, we downloaded and screened transcriptome data of study-compliant NSCLC samples and paracancerous control samples. First, we used single-cell RNA-sequencing (scRNA-seq) data to identify cell subsets in different differentiation states by trajectory analysis and to identify important NSCLC cell differentiation-related genes (CDRGs). Second, we explored the biological functions of CDRGs and found that they are involved in tumor immune regulation and immune responses. We predict and verify the different differentiation trajectories and tumor-promoting mechanisms of NSCLC cells, providing an important data basis for further modeling to predict tumor immunotherapy response and patient survival.

2. Manuscript Preparation

2.1. Data download and collection

This study mainly downloads data from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The retrieval conditions are set as follows: (1) Keywords: lung cancer (set the pathological type as non-small cell lung cancer), (2) File options: the data type is "single-cell sequencing", and the data category is "transcriptome analysis". Download the single-cell RNA sequencing information of primary lung cancer patients uploaded as of June 1, 2020, and download 16 case samples, of which 11 samples are primary non-small cell lung cancer, and 5 sample is from adjacent normal lung tissue as a control group, a total of 4320 cells were detected, using Smart-seq2 technology, and the sequencing platform was Illumina Next Seq 500.

2.2. Identify highly variable genes

The feature subsets showing high inter-cell variation in the dataset were selected by the "Find Variable Features" function of the "Seurat" function package to select highly variable genes (HVGs), and the relationship between the mean and variance of expression was calculated. Using the best method: linearly fit the log (variance) and log (mean) with loess, standardize the mean of the detected target gene and the variance of the expected target gene, and calculate the variance of the target gene with the maximum standardized gene expression.

2.3. Principal Component (PCA) Analysis and Nonlinear Dimensionality Reduction (T-SNE)

To reduce the effect of tumor-to-tumor variability on the comprehensive analysis of tumor cells, we refocused the data within each tumor individually so that the mean value for each gene in each tumor cell was zero. The variance matrix for PCA was generated using the method outlined by Shalek et al [12]. To reduce the weight of less reliable "missing" values in the data. The purpose of the T-distributed Stochastic Neighbor Embedding (T-SNE) algorithm is to understand the diversity of data and perform low-dimensional clustering. Calling the T-SNE function to map the data in the high-dimensional space to the low-dimensional space based on retaining the local features of the data set is one of the better methods for data dimensionality reduction and visualization. The disadvantage is that it takes up a lot of memory and calculates the running time of the function for a long time.

2.4. Orbital Analysis and CDRG Identification

Differentiation trajectories of tumor cells in NCSLC-scRNAseq samples were described using the Monocle 2 technology [13]. Individual cells are projected into this space and arranged in trajectories with branch points. Cells in the same branch are generally considered to be in the same state of differentiation, while cells located in different branches are considered to have different cellular differentiation characteristics. Then, gene differential expression analysis was performed between each branch, and the significantly differentially expressed genes were defined as branch-dependent genes, and were captured and marked by the software. These differentially expressed marker genes located in different clades were

ISSN 2706-6819 Vol.5, Issue 10: 81-87, DOI: 10.25236/IJFM.2023.051013

defined as cell differentiation-related genes (CDRGs).

2.5. Gene Set Enrichment Analysis (GSEA)

GSEA (http://software.broadinstitute.org/gsea/index.jsp) is a gene probe enrichment assay based on the evaluation of various levels of gene probes from microarray data. GSEA generates an ordered list of all genes based on the correlation of selected gene expression and performs KEGG signaling pathway analysis and GO biological function enrichment analysis, which is used to identify the relevant molecular mechanisms of NSCLC cells in different differentiation states.

2.6. Statistical analysis

All data of the study were conducted using R package software (version 3.6.1) and Perl scripting tools (version 5.30.0). The mean and standard deviation represent continuous variables, while the frequency and percentage represent categorical variables. The results with the default $P < 0.05$ were statistically significant and could be included in the next analysis.

3. Results

3.1. Identification of related genes based on single-cell sequencing data

Following quality control criteria and normalization of NSCLC-scRNAseq data, 191 low-quality cells were excluded and 4129 cells from NSCLC cores were included in the analysis, including a total of 14901 corresponding genes. The ratio of ribosomal genes and the number of genes in the cells are calculated separately by the "Seurat" package of the R software, "nFeature RNA" is the number of detected genes whose expression level is greater than 0, and the count is the sum of the expression levels of all genes. The number of genes, the number of cells, and the proportion of genes were marked and their distribution frequencies were counted (Figure 1).

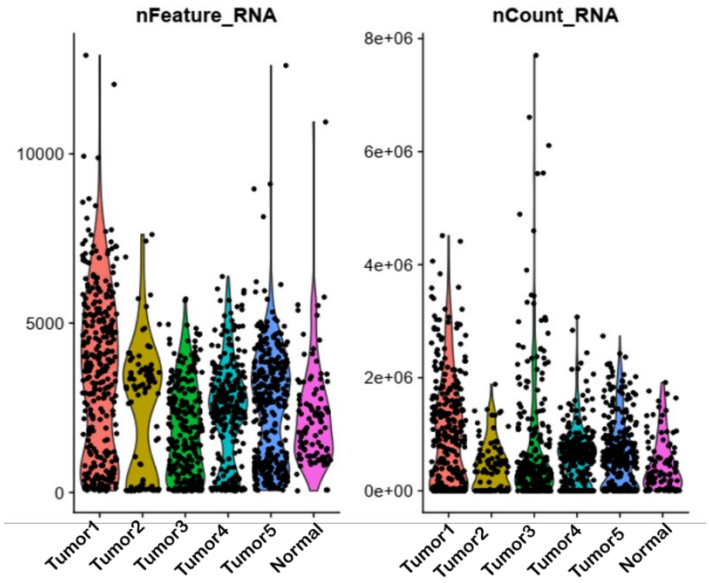

Figure 1: The number of genes and the proportion of gene expression markers and their distribution frequencies from 6 samples.

3.2. Data normalization

For standard selection, we filtered cells based on gene count and mitochondrial ratio, filtering for the lowest gene count, and selecting cells with a ratio greater than 50% of the characteristic standard count. In our selected dataset, no mitochondrial highly expressed genes were detected, indicating that all cells were in good condition, and all cells were reserved for further analysis. The results showed that the number of detected genes was significantly correlated with the sequencing depth, with a Pearson correlation coefficient of 0.63 (Figure 2).

ISSN 2706-6819 Vol.5, Issue 10: 81-87, DOI: 10.25236/IJFM.2023.051013

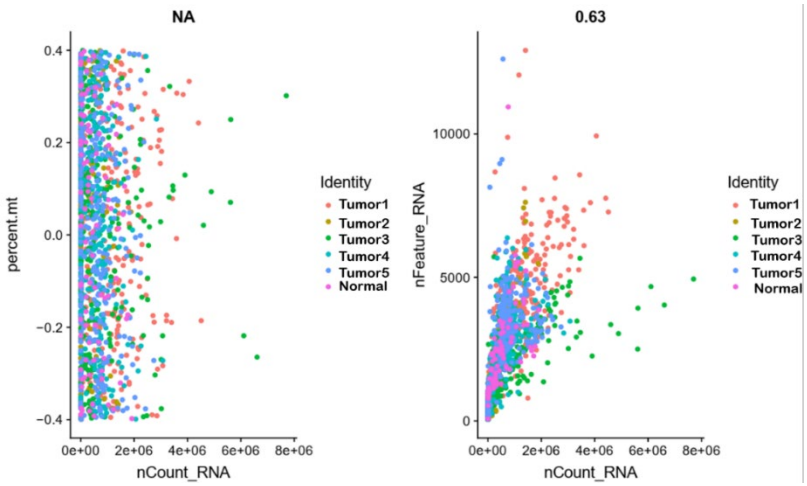

Figure 2: Number of genes significantly correlated with sequencing depth.

3.3. Identify and screen hypervariable genes

ANOVA plot showing 1,500 highly variable genes out of 14901 genes from NSCLC samples. Red dots represent highly variable genes and black dots represent immutable genes. CCL4, GZMK, TNFRSF4, KLRG1, GNLY, C12orf5, FGFBP2, MYO1G, SLC25A, and 14APEH with gene names marked in the figure are the 10 genes with the highest degree of variation (Figure 3).

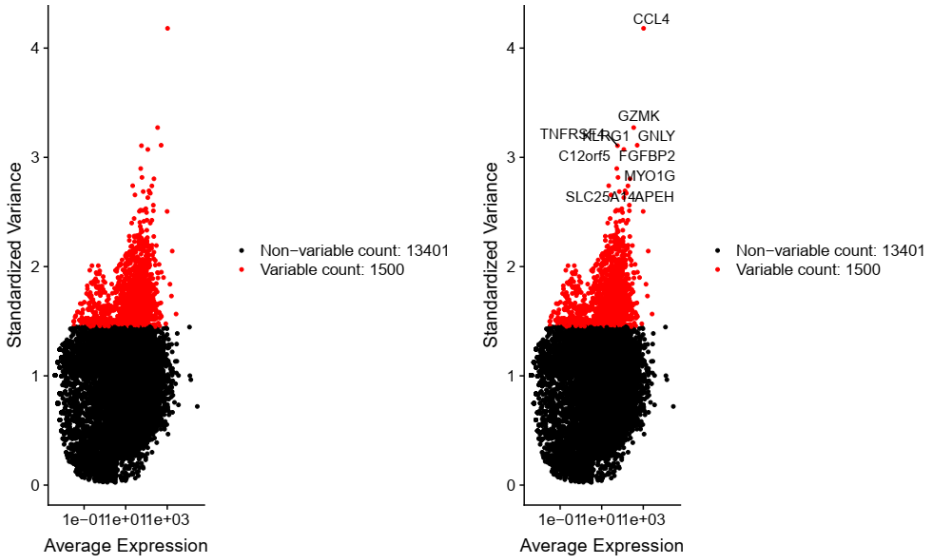

Figure 3: Variance plot of highly variable gene signatures between cells.

3.4. PCA principal component analysis and T-SNE dimensionality reduction

Principal component analysis (PCA) was used to determine available dimensions and screen for related genes. PCA results did not show a clear separation between cells in NSCLC (Figure 4A). We selected the top 20 principal components (PCs) with $P<0.05$ for subsequent analysis (Figure 4B). Applying the T-SNE algorithm to the dimensionality reduction of 20 PCs successfully classified 3 cell clusters, and the clustering results are shown in Figure 4C. Using the Wilcox method, we set the screening index to logFCfilter to 0.5 and adjPvalFilter to 0.05 to find significantly high-expressed genes in each cluster and screened out 9876 differentially expressed genes. Then we performed cluster analysis on the top differential genes. The clustering results showed that 10 differential marker genes could be clustered in different clusters. The colors from purple to yellow indicate gene expression levels from low to high (Figure 4D).

ISSN 2706-6819 Vol.5, Issue 10: 81-87, DOI: 10.25236/IJFM.2023.051013

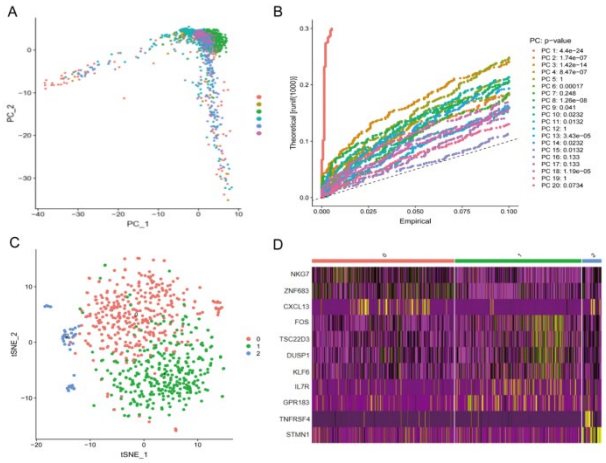

Figure 4: Related genes after PCA and T-SNE dimensionality reduction. (A) PCA does not show a clear separation of cells in NSCLC. (B) PCA identified 20 PCs with, an estimated P < 0.05. (C) T-SNE dimensionality reduction analysis classified 3 cell clusters. (D) The top 10 marker genes for each cell cluster are shown in the heatmap.

3.5. Tumor differentiation trajectory analysis and determination of biological functions of NSCLC differentiation-related genes

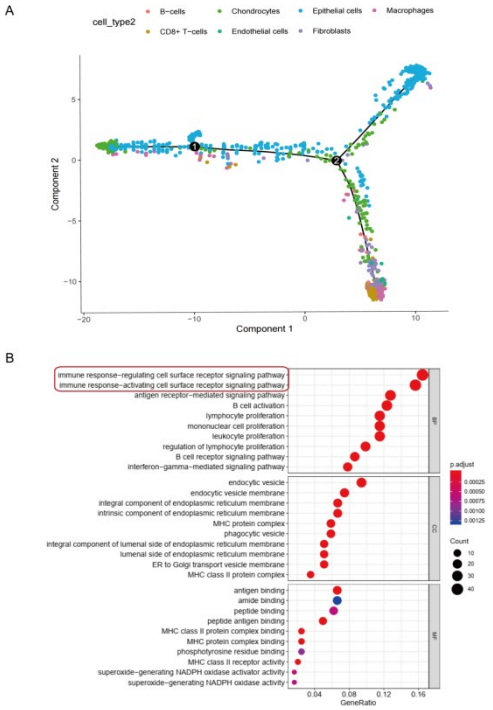

Figure 5: Cell annotation, trajectory analysis, and GSEA analysis of three GBM cell subsets with different differentiation patterns. (A) Trajectory analysis showing three subpopulations of NSCLC cells with distinct differentiation patterns. NSCLC CSCs were mainly distributed in roots, whereas NSCLC cells were distributed in three branches. (B) GSEA-GO analysis showed that the three subtypes of NSCLC cell subsets were significantly associated with immune regulation and immune response function.

Using Monocle 2 technology to analyze the cell differentiation trajectory, it can be seen that the tumor cells differentiate into 3 branches, and each branch has different NSCLC immune-related cells. Among them, branch I distributed 410 cells, branch II distributed 404 cells, and branch III distributed 438 cells (Figure 5A). The three branches are defined by type I, II, and III cell subsets, respectively. Gene difference analysis obtained 275 type I CDRGs, 191 type II CDRGs, and 198 type III CDRGs, and the differences in the degree of differentiation of the three types of cell subsets were statistically significant.

ISSN 2706-6819 Vol.5, Issue 10: 81-87, DOI: 10.25236/IJFM.2023.051013

Finally, GSEA functional enrichment analysis found that type I CDRGs were significantly associated with immune response modulation, while type II and III CDRGs were significantly associated with immune response pathways (Figure 5B).

4. Discussion

NSCLC is the most common tumor, and the number of new cases and deaths still ranks first among malignant tumors. This tumor has significant tumor heterogeneity in the process of diagnosis and treatment [14]. Heterogeneity is characteristic of many cancers, such as lung cancer, and is associated with clinical progression and an important driver of drug resistance, so the analysis of different cell species in tumors is extremely important to reveal the mechanism of drug resistance. Single-cell sequencing technology enables specific analysis of cell populations at the single-cell level. The process mainly includes single-cell isolation, cell lysis and genomic DNA acquisition, whole genome amplification, sequencing, and data analysis. Single-cell sequencing includes single-cell whole genome sequencing and single-cell transcriptome sequencing, which explore the impact of highly heterogeneous cells on diseases from the genomic and transcriptome levels, respectively, and identify the main cell subsets that affect them, to provide a basis for the diagnosis and treatment of diseases. During the development of NSCLC, cells are constantly differentiated and mutated into different cell subsets to enhance immune suppression and immune evasion. As intratumor heterogeneity is increasingly recognized as one of the main reasons for tumor therapy resistance, there is an urgent need to develop new technologies to study cellular heterogeneity in NSCLC [15] deeply. However, up to now, studies on the cellular heterogeneity, microenvironmental heterogeneity, and tumor immune heterogeneity of NSCLC are limited. Therefore, we preliminarily explored the heterogeneity results of NSCLC in genetic testing from the level of single-cell RNA, to provide a data basis for the study of tumor heterogeneity in NSCLC.

With the advent of the era of lung cancer immunotherapy, it is increasingly recognized that metabolic changes in cancer cells can affect immune cell function and lead to tumor immune evasion. It has a certain impact on the effect of immunotherapy [16]. Immune cells in the tumor stroma sometimes colonize an environment with different cell subsets and nutrients as they tour the body, and the cross-talk between tumor cells and immune cells ultimately results in an environment that promotes tumor growth and metastasis, In medicine, it is called the tumor immune microenvironment [17, 18]. A large number of studies have shown that the heterogeneity of the tumor immune microenvironment affects the immunotherapy effect of tumor patients from many aspects such as genetics and immunity [19]. An in-depth understanding of the heterogeneity of this environment will facilitate the development of therapeutic approaches that simultaneously target multiple components of the immune microenvironment, thereby increasing the likelihood of good clinical outcomes [20].

Single-cell transcriptome sequencing enables a dynamic representation of gene lineage and heterogeneity to better define the cell types examined [21]. Almost all studies of predictive biomarkers associated with clinical prognosis in tumors are based on gene-level analysis of a single biopsy sample [22]. Based on single-cell scRNA-seq sequencing technology, this study compared and analyzed different cell subsets in tumor samples, and predicted the differentiation trajectory of tumor cells and their differentially expressed cell differentiation-related genes. The early stages of occurrence have already emerged. In this study, we identified 3 cell clusters from 11 NSCLC samples, and based on cell trajectory analysis, NSCLC cells were projected into three subpopulations with significantly different differentiation characteristics. Screening identified subpopulation-dependent Cell Differentiation-Related Genes (CDRGs). Through GSEA-GO biological function correlation analysis, we found that this differentiation model was significantly associated with tumor immune regulation and immune response, implying an intrinsic correlation between NSCLC cell differentiation and intratumoral immune and metabolic biology. Of course, the current research still has certain limitations. On the one hand, the patient details obtained by downloading are not complete enough, and some clinical parameters, such as tumor imaging results, medical records medical history, and details of surgical records, cannot be downloaded, so the nomogram cannot be input. , on the other hand, has a limited number of cases available for download. To make this study more clinically meaningful, the predictive model needs to be further validated in future large-scale cohorts.

In summary, we predicted NSCLC cells with different differentiation characteristics based on the scRNA-seq data of the GEO database and then performed differential expression analysis to find cell differentiation-related genes (CDRGs). Its biological functions and metabolic pathways are involved. This study highlights the unique cellular differentiation trajectories of NSCLC cells and their important role in predicting clinical outcomes and tumor immunotherapy response in predicting clinical outcomes and tumor immunotherapy responses in lung cancer patients.

ISSN 2706-6819 Vol.5, Issue 10: 81-87, DOI: 10.25236/IJFM.2023.051013

5. Conclusion

Our study identified NSCLC cells with distinct differentiation characteristics based on single-cell sequencing data from GEO, emphasizing the important role of cell differentiation in predicting the clinical outcome of NSCLC patients and their potential response to immunotherapy.

Acknowledgments

The results shown here are in whole or part based upon data generated by the GEO database: https://www.ncbi.nlm.nih.gov/geo/.

References

[1] Torre LA, Siegel RL, Jemal A. Lung cancer statistics. Adv Exp Med Biol. 2016, 893(1):1–19.
[2] Seow WJ, Matsuo K, Hsiung CA. Association between GWAS-identified lung adenocarcinoma susceptibility loci and EGFR mutations in never-smoking Asian women, and comparison with findings from Western populations. Hum Mol Genet, 2017, 26(1): 454–465.
[3] Goveia J, Rohlenova K, Taverna F. An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. Cancer Cell. 2020, 37(1): 21-36.
[4] Kron A, Scheffler M, Heydt C. Genetic Heterogeneity of MET-Aberrant NSCLC and Its Impact on the Outcome of Immunotherapy. J Thorac Oncol. 2021, 16(4): 572-582.
[5] Jia Q, Wu W, Wang Y. Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. Nat Commun. 2018, 9(1): 53-61.
[6] Sui H, Ma N, Wang Y. Anti-PD-1/PD-L1 Therapy for Non-Small-Cell Lung Cancer: Toward Personalized Medicine and Combination Strategies. J Immunol Res. 2018, 8(1): 20-69.
[7] Klein AM, Mazutis L, Akartuna I. Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell, 2015, 161(3): 1187–1201.
[8] Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. Nat Biotechnol, 2016, 34(4): 1145–1160.
[9] Zheng GX, Terry JM, Belgrader P. Massively parallel digital transcriptional profiling of single cells. Cancer Manag Res, 2019, 11(1): 7197-7210.
[10] Azizi E, Carr AJ, Plitas G. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell, 2018, 174(1): 1293–1308.
[11] Peng J, Sun BF, Chen CY. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res, 2019, 29(1): 725–738.
[12] Shalek AK, Satija R, Shuga J. Single-cell RNA-seq reveals dynamic paracrine control cellular variation. Nature. 2014, 510(7): 363-369.
[13] Qiu X, Mao Q, Tang Y. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods, 2017, 14(2): 979–982.
[14] Jonna S, Subramaniam DS. Molecular diagnostics and targeted therapies in non-small cell lung cancer (NSCLC): an update. Discov Med. 2019, 27(148): 167-170.
[15] Pe'er D, Ogawa S, Elhanani O. Tumor heterogeneity. Cancer Cell. 2021, 39(8): 1015-1017.
[16] Ren X, Zhang L, Zhang Y. Insights Gained from Single-Cell Analysis of Immune Cells in the Tumor Microenvironment. Annu Rev Immunol. 2021 Apr 26;39:583-609.
[17] Sun YF, Wu L, Liu SP. Dissecting spatial heterogeneity and the immune-evasion mechanism of CTCs by single-cell RNA-seq in hepatocellular carcinoma. Nat Commun. 2021, 12(1): 40-91.
[18] Cooper LA, Demicco EG, Saltz JH. PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective. J Pathol. 2018, 244(5): 512-524.
[19] Chen YP, Lv JW, Mao YP. Unraveling tumor microenvironment heterogeneity in nasopharyngeal carcinoma identifies biologically distinct immune subtypes predicting prognosis and immunotherapy responses. Mol Cancer. 2021, 20(1): 14-31.
[20] Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer. 2019, 19(3): 133-150.
[21] Gao S. Data Analysis in Single-Cell Transcriptome Sequencing. Methods Mol Biol. 2018, 1754(1): 311-326.
[22] Stepan H, Hund M, Andraczek T. Combining Biomarkers to Predict Pregnancy Complications and Redefine Preeclampsia: The Angiogenic-Placental Syndrome. Hypertension. 2020 Apr;75(4):918-926.