Immune Infiltration Analysis by TIMER, quanTIseq, and Cibersort

Chenling Zhou [1], Shiyu Fan [2, *]

[1] School of Arts and Sciences, Rutgers University, New Brunswick, USA, 08901

[2] School of Science, University of Auckland, Auckland, New Zealand, 1142

* Corresponding Author Email: sfan289@aucklanduni.ac.nz

Abstract. Immune Infiltration analysis is of great significance in cancer research and can provide insight into the activity of the immune system in different disease states. The expression profile analysis of high-throughput sequencing provides a powerful tool for immune cell research, but different methods may lead to different results. This study used a dataset published by Ma et al in August 2023, which included peripheral blood mononuclear cells (PBMCs) from 19 patients with non-small cell lung cancer (NSCLC) and four healthy human donors. This papper used three immune cell estimation methods (TIMER, quanTIseq, and Cibersort) for analysis, and found differential immune cell proportions from the three methods. To accurately determine the proportion of immune cells, this passage weighted the results of the three methods. The results show that the weighted average method can more comprehensively reveal the presence and distribution of different immune cell types. This provides important insights for immunological research and cancer treatment, underscoring the importance of integrating multiple analytical approaches.

Keywords: Immune Infiltration Analysis, High-throughput Sequencing, TIMER, QuanTIseq, Cibersort.

1. Introduction

Immune Infiltration analysis is a method used to study the presence and composition of immune cells in biological samples and is often widely used in cancer research. The main aim is to understand the extent of immune system activity in a particular disease state and its potential impact on the disease. In this paper, we will use steps such as data acquisition, data preprocessing, identification of immune cell markers, and estimation of immune cell infiltration, data analysis, visualization and interpretation of results to study lung cancer cells [1].

Our analysis and discussions in this paper mainly depends on the Immuno-Oncology Biological Research (IOBR) tool. “IOBR offers batch analyses of these signatures and their correlations with clinical phenotypes, long non-coding RNA (lncRNA) profiling, genomic characteristics, and signatures generated from single-cell RNA sequencing (scRNA-seq) data in different cancer settings. Additionally, IOBR integrates multiple existing microenvironmental deconvolution methodologies and signature construction tools for convenient comparison and selection. Collectively, IOBR is a user-friendly tool for leveraging multiomics data to facilitate immuno-oncology exploration and to unveil tumor-immune interactions and accelerating precision immunotherapy” [2]. IOBR brings together various pre-existing methods and tools. CIBERSORT is adept at extracting meaningful cell composition information from complex gene expression datasets. TIMER covers a diverse array of cancer types, permitting an expansive range of immune infiltration analyses across different datasets. Unlike other tools that only offer relative proportions, QuanTIseq provides estimates of the absolute numbers of each immune cell type. This integrated approach would help researchers to save time and ensure to operate on the same platform.

Immune Infiltration analysis provides critical information about the types and numbers of immune cells present, contributing to deep studies of immune-related disease mechanisms, while also providing useful information for personalized treatment. Different studies and experiments may require the use of different tools. In this study, we used three different methods, namely TIMER, quanTIseq and Cibersort. By weighted averaging their results, we aim to more accurately determine

the proportion of each immune cell type in the cell. There are errors in single methods, so by weighting their results to average, we aim to more accurately determine the proportion of each immune cell type in the cell. The significance of this study is to delve deeper into the intricate relationship between the immune system and lung cancer, providing the basis for a deeper understanding of the pathogenesis of the disease and paving the way for the development of personalized treatments.

2. Data Description

This set of data was compiled by Ma W, Wei S, Tian EC et al., named "Multicolor spectral flow cytometry evaluation of peripheral blood mononuclear cells from NSCLC, 3 August 2023 patients" [3]. Expression profiling by high-throughput sequencing, a powerful technique widely used in molecular biology and genomics, was used in the study. It provides a comprehensive analysis of gene expression patterns within biological samples through next-generation sequencing technology, providing intensive and quantitative insights into the transcriptome of an organism, specific tissue, or cell type.

The main subjects of the dataset were peripheral blood mononuclear cells (PBMCs) from 19 patients with non-small cell lung cancer (NSCLC) before and after immune checkpoint inhibitor (ICI) treatment and from four healthy human donors. The study used multicolor spectral flow cytometry to simultaneously monitor 24 immune cell markers during treatment. This set of data mainly contains 12 types of cells (which will be expressed in the form of A1 and B2 in the later experiment) and their gene types, mainly B cell, T cell, NK cell, etc.

Can be in the following link for more detailed information about the data set: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE235048

3. Method

3.1. TIMER

TIMER is abbreviation for Tumor Immune Estimation Resource. As a tumor immune estimation tool and database, TIMER is a comprehensive resource for systematic analysis of immune infiltration in different cancer types [4].

After we collect the required data, we will remove some of the highly expressed genes from the data, considering that they lack stability. After removing the part of data set, we call itG_{0f}. Then, we performed constrained least squares fit on each sample to predict the relative abundance of each of the six immune components. Least squares method will show the best functional match of the data by minimizing the sum of squares of the errors. We treat each given sample as a mixture of six immune cell types, so our goal is to find a set of positive coefficients f that minimizes the total squared variance. In the filtered sample, let g belong to our processed dataset G0f. In addition, we adjusted the immune cell type r (r = 1, 2... The expression level of gene g in 6) is represented byX_r^g. From this, we can get a constrained linear regression formula:

$$f = argmin_{\forall_r:f_r>0} \sum_{g\in\{G_{0f}\}} \left(Y^g - \sum_{r=1}^{6} f_r X_r^g\right)^2 \quad (1)$$

Then we need to estimate the total number of white blood cells based on DNA methylation. Suppose "T_{ik}" represents the beta value of probe *k* in tumor sample *i*, which is a measure of DNA methylation levels; "B_k" represents the average beta value of each probe in normal tissue samples (buffy coat samples) for reference; "T_k" represents the smallest beta value of the BC probe observed in all tumor samples. These values could theoretically reflect the ground state of methylation levels in the purest tumors; "f_B" represents the proportion of buffy coat (leukocyte) components in the sample, a parameter used to estimate the proportion of normal cells. Based on these assumptions, we can derive two linear equations that may be used to analyze DNA methylation data in tumor samples.

$$T_{ik} = B_k f_B + T_k(1 - f_B) \tag{2}$$

$$f_B = (T_{ik} - T_k)/(B_k - T_k) \tag{3}$$

TIMER has the following advantages:

First, the TIMER supports many different types of tumors, which allows researchers to compare and analyze between different cancer types to understand differences in immune cell infiltration and immune-related gene expression. At the same time, TIMER allows users to estimate the abundance of different types of immune cells in tumor tissue, which has important implications for studying the tumor immune environment and predicting patient survival. Researchers can obtain information about the number of immune cells in the tumor microenvironment. In addition, TIMER allows for survival analysis of tumor patients, which helps to determine the association between immune factors and patient survival, providing important information for clinical studies.

3.2. QuanTIseq

In quanTIseq, we will calculate the TIL10 signature matrix. Since highly expressed genes will affect the deconvolution results, we also first remove highly expressed feature genes from the data. However, we made a judgment based on the theory in cancer cell Line Encyclopedia in quanTlseq. We defined the genes with an average log2 expression greater than 7 as highly expressed feature genes, which means genes expressing more than 700 TPM, and filtered them out. After the data is preprocessed, a microarray assembly of the CIBERSORT LM22 signature matrix is constructed with the preprocessed data [5].

In order to solve the problem of underestimation of T_{reg} cells and CD4 T cells due to multicollinearity, we adopted a heuristic strategy to solve the problem. First, by estimating the Treg cell fractions F_{reg}^1 andF_{reg}^2, and then taking their average asF_{reg}. Then take the larger value between the difference of F_{CD4}^1 and Freg and 0 as the final value ofF_{CD4}.

$$F_{reg} = mean\left(F_{reg}^1, F_{reg}^2\right) \tag{4}$$

$$F_{CD4} = max\left(F_{CD4}^1 - F_{reg}, 0\right) \tag{5}$$

In quanTIseq, we use IHCount's computational workflow to help us quantify the number of total cells and tumor-infiltrating immune cells in the image. IHCoun is a generic workflow for quantitative analysis of multiple ihc images. It helps to preprocess images and measure cell counts. Image processing is performed using several intensity-based modules to identify and quantify positively stained cells, nuclei, and tissue regions. We will then classify the pixels of the positive stained cell and the pixels of the nucleus, forming two sets to facilitate subsequent processing. Both collections export a multi-channel TIF (32-bit float), which represents a TIF image file containing multiple image channels that can be used to store high-quality image data.

Meanwhile, we also use the Cellprofiler pipeline in quanTIseq. The CellProfiler pipeline is a workflow for defining and executing image analysis tasks. We use the CellProfiler pipeline to create a series of image processing and analysis steps that can be performed automatically in a specific order to enable the processing and analysis of cell image data. We can extract quantitative data from the images, including cell number, size, shape, color and other characteristics, which is very important for biological research.

In the process of forming the picture, we will use the t-SNE algorithm to exclude the fraction of uncharacterized cells, and then estimate the immune cell fraction. T-SNE is a nonlinear dimensionality reduction machine learning algorithm for exploring high-dimensional data. It maps multidimensional data to two or more dimensions suitable for human observation. It is based on the probability distribution of random walks on the neighborhood graph to find the structure within the data.

3.3. Cibersort

Cibersort mainly uses GEP deconvolution, and its core idea is to use generalized eigenvalue problems to model the reconstruction process of images. The generalized eigenvalue problem is A matrix problem, usually of the form Av = λBv, where A and B are matrices, v is an eigenvector, and λ is an eigenvalue. GEP deconvolution can capture complex relationships between images, rather than just linear ones, and therefore performs well in cases involving nonlinear image changes. We can use the conditional number as an inherent matrix property to enhance the robustness of the gene feature matrix. Robustness represents the consistency and stability of this feature matrix for gene expression data, especially in the face of different experimental conditions, different data sets or different gene expression measurement platforms. Robustness represents the ability of the feature matrix to maintain relatively consistent performance in different situations.

After establishing the matrix, we will evaluate the stability and consistency of the model through the kappa value. We can obtain the kappa value through the following formula, where P_O is the proportion of the actual number of samples consistent with the prediction. P_e Is the expectation of random classification accuracy, calculated by randomly assigning a sample to a class and then calculating the expected value of the observer's classification accuracy?

$$\kappa = \frac{P_O - P_e}{1 - P_e} \tag{6}$$

Kappa value less than 0 indicates that the consistency of model prediction is worse than random selection. The Kappa value approaching 0 indicates that the performance of the model is similar to that of random selection. A Kappa value close to 1 indicates that the performance of the model is very good, and its predictions are very consistent with actual observations.

After obtaining a model with good performance, we will build the signature matrix with LM22. "The process of building a signature matrix represents a type of 'filter method', a preprocessing step that removes irrelevant features before application of a specific machine learning approach or prediction algorithm21. Specifically, the use of a signature matrix facilitates (i) faster computational running time owing to the elimi- nation of genes with uniform expression levels across the cell types of interest (for example, housekeeping genes and unexpressed genes) and (ii) a greater signal-to-noise ratio by preselecting reference profiles that have maximal discriminatory power (as measured by condition number)" [6]. The process of constructing the signature matrix can help us estimate the signature gene expression matrix for the proportion of white blood cells in bulk RNA.

The LM22 signature matrix is a gene signature matrix used to describe the relative expression levels of immune cell types in tumor tissues. It is one of the tools used for immune cell type analysis and is commonly used in studies of tumor immunology, immunotherapy and tumor microenvironment. It contains gene expression signatures for 22 different immune cell subtypes. These immune cell subtypes include various T cells, B cells, natural killer cells, macrophages, and other immune-related cell types. Generation of the LM22 signature matrix is accomplished by analysis of known immune cell-specific marker genes associated with different immune cell types. It can be used to evaluate the relative abundance of different immune cell types in tumor tissues to understand the distribution and activity of immune cells in the tumor microenvironment. Used to evaluate the effect of immunotherapy on immune cells in tumor tissue to monitor the effect of treatment and predict patient response; It is used for tumor classification and prognostic assessment, as the presence of different immune cell types can be associated with tumor pathologic features and patient prognosis.

Cibersort can infer the relative presence ratio of different cell types from gene expression data, allowing researchers to better understand the cellular composition of complex tissues. In addition, Cibersort performed well when dealing with noisy data and unknown mixture content. This means it is able to cope with challenges from real samples, even if the sample contains an unknown mixture or there are measurement errors.

4. Results

In this study, we studied a set of lung cancer data for 2023, aiming to conduct an in-depth expression profile analysis. The dataset first undergoes the necessary pre-processing steps to ensure the accuracy and visualization of the data [7]. The steps include:

4.1. Data Selection

We first selected the most recent lung cancer data set to be published in 2023 to ensure we were basing our study on the most up-to-date information.

4.2. Data Preprocessing

Data preprocessing is a key step to ensure the quality of data. This includes processing missing values, outliers, and normalized data for subsequent analysis.

4.3. Data Structuring

We reorganized the cell names in the data into column names and the gene names into row names for better analysis and graphical presentation. This adjustment of the data structure helps to improve the readability and clarity of the data in order to better understand the analysis results.

4.4. Deconvolution analysis

We used a variety of deconvolution methods, including TIMER, quanTIseq, and Cibersort, to identify and quantify the expression of different cell types in lung cancer tissue. This analysis helps us understand the presence and activity of immune cells and other cell types in the tumor microenvironment.

4.5. Signature expression Analysis

We further performed signature expression analysis on the data to reveal biological features and pathways associated with lung cancer. This helps us understand the underlying biological mechanisms and therapeutic targets.

In the process of TIMER backwinder analysis, we have encountered an important problem, that is, the standardization of raw data is different. Specifically, the standardized amount of each cell type is not consistent, which can be influenced by a variety of factors. To overcome this challenge, we took an additional normalization step and normalized the data generated by the TIMER again. Standardized practices ensure that data is comparable across different cell types and samples. By re-standardizing, we can compare the amount of expression between different cell types and ensure that we take this into account when analyzing and interpreting the data.

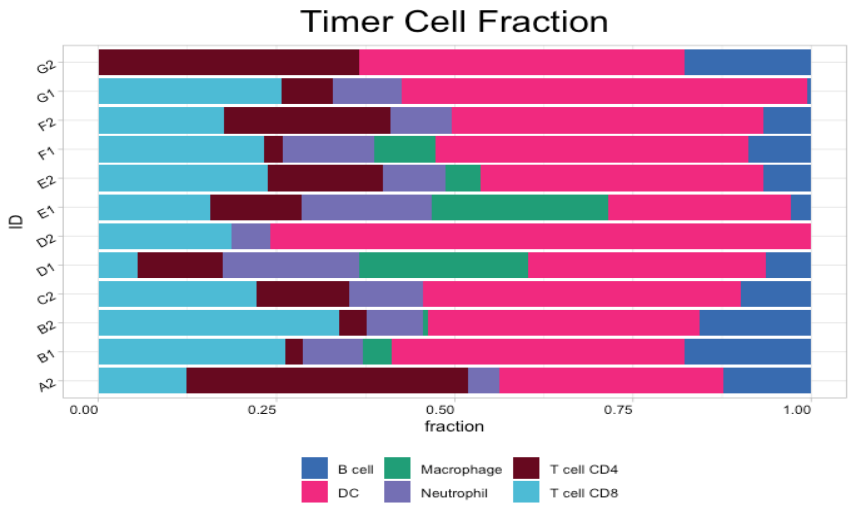

Figure 1. Timer Cell Fraction.

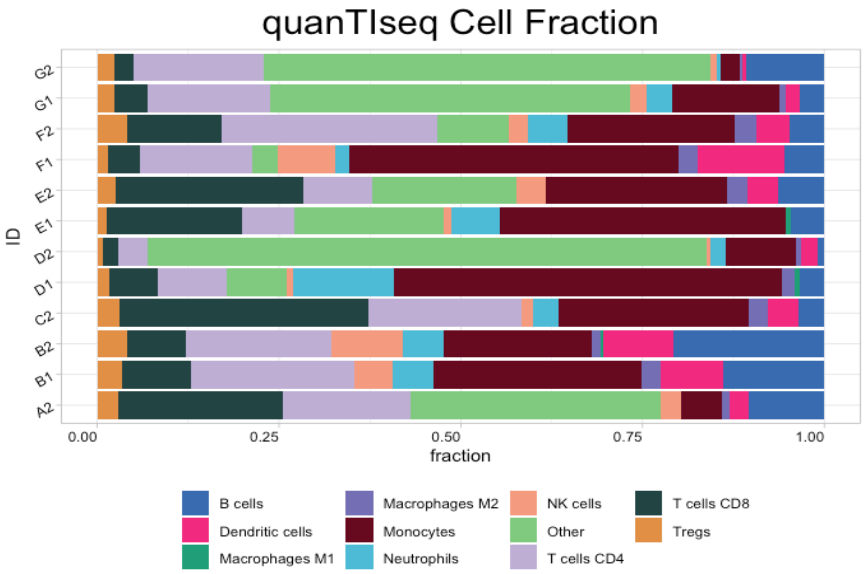

Figure 2. QuanTIseq Cell Fraction.

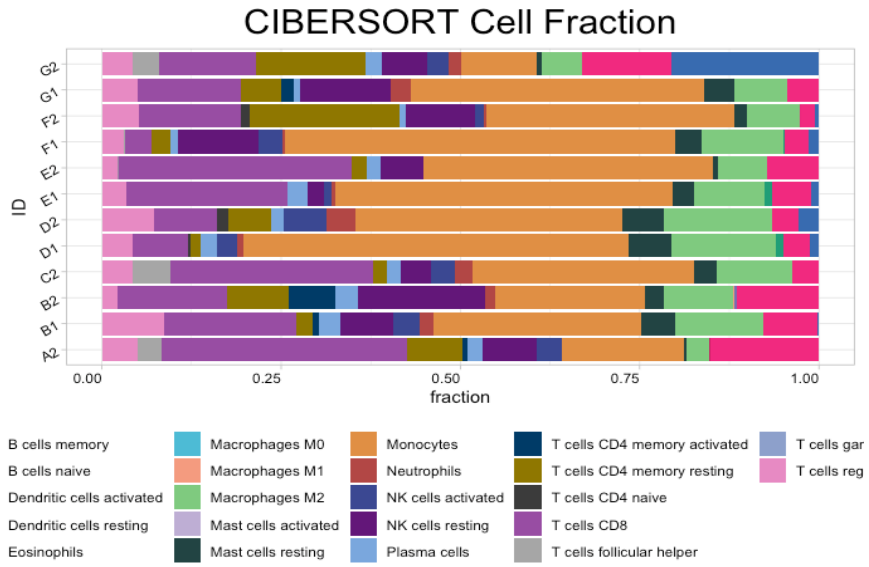

Figure 3. CIBERSORT Cell Fraction.

Through the above steps, we obtained three charts, shown in Figure 1 to 3, each representing an in-depth analysis of immune cell expression in lung cancer tissue using different analytical methods (TIMER, quanTIseq, and Cibersort). However, one challenge we encountered when comparing the results of different methods was the difference in the types of immune cells identified by each method. This may be due to different algorithms, data processing strategies, and standardized methods. To solve this problem, we turned to the immune cell type classification method provided in the research article "sc-ImmuCC: hierarchical annotation for Immune Cell types in single-cell RNA-seq". This article provides a hierarchical classification system for immune cell types that can unify the naming of different cell types. This classification method allows us to map the results of each assay into a common framework for naming immune cell types, thus achieving consistency and comparability of results.

In the analysis of immune cell types, it is particularly necessary to make large adjustments to B cells, macrophages, and T cells. “According to the differentiation lineage of the immune cells, we divided the annotation process of immune cells into three layers. The first layer consists of nine major immune cell types: T cells, B cells, monocytes, macrophages.....The second layer of cells is the subtype of the first layer of cells, mainly including B cell subtypes: naïve B cells, memory B cells

and plasma cells; T cell subtypes: CD4 T cells and CD8 T cells; NK subtypes: NK_bright and NK_dim; and macrophage subtypes: M1 macrophage and M2 macrophage. There are a total of 16 cell subtypes in the second layer. The third layer is a more specific subtype classification for the CD4 T cells and CD8 T cells in the second layer" [8]. We took the following strategy to make these adjustments: We merged the B cells memory and B cells naive from Cibersort and merged them into B cells. This helps to consider the presence and function of B cells more fully; We pooled Macrophages in Cibersort and quanTIseq, including types M0, M1 and M2. This combination helps to consider the immune properties of macrophages and their role in lung cancer. We made a finer classification of the T cells in Cibersort and quanTIseq, classifying them according to markers of CD4 and CD8. This classification helps distinguish between different types of T cells to gain a more complete understanding of their function and importance in lung cancer research. With these adjustments, we were able to analyze different immune cell types more accurately, providing a more detailed and comprehensive insight into the composition of immune cells in lung cancer tissue. These strategies help ensure the accuracy and interpretability of the data, further improving the depth and dimension of the research [9].

Next, we used a weighted average of the results of the three methods to generate the final immune cell distribution table, shown in Figure 4. This table includes the following six immune cell types: Dendritic cells, Neutrophil, CD4-positive T cells, Macrophage, other immune cell types, and CD8 positive T cells. Through this comprehensive approach, we were able to gain a more comprehensive understanding of the relative distribution and expression levels of these six key immune cell types in lung cancer tissues, providing a powerful tool and data foundation for in-depth studies of tumor immunology and disease biology [10]. This comprehensive table reflects the common insights of different analytical approaches, helping to gain a more global perspective to support the development of lung cancer research and treatment strategies.

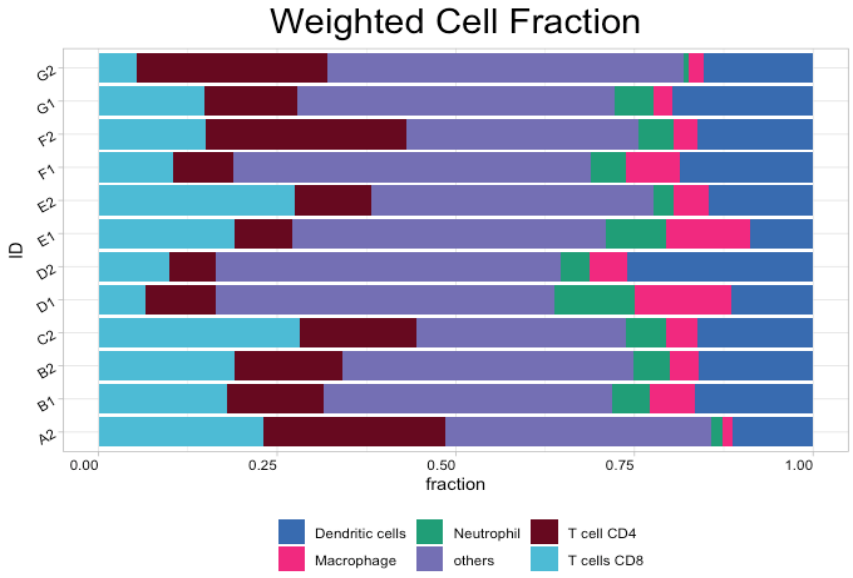

Figure 4. Weighted Cell Fraction.

5. Conclusions

To sum up, we conducted an in-depth analysis of immune cells in lung cancer tissue in this study using three different immune cell estimation methods (TIMER, quanTIseq, and Cibersort). We faced differences in immune cell type recognition between different approaches, but through careful classification and adjustment, we managed to fuse them together for a more comprehensive result.

By weighted averaging the results of these three methods, we constructed a final distribution table containing six key immune cell types: dendritic cells, neutrophils, CD4-positive T cells, CD8A-

positive T cells, macrophages, and other immune cell types. This comprehensive table provides detailed insights into the relative presence and expression levels of immune cells in lung cancer tissue.

Our study provides important information and insights into the lung cancer immune microenvironment. By standardizing and classifying, we overcome the differences between different approaches to provide more reliable and comparable data for immunological studies of lung cancer. This will help reveal the key role of immune cells in the development and treatment of lung cancer, providing strong support for the development of personalized medicine and treatment strategies. Our study highlights the complementarity of multiple analytical approaches, highlighting the importance of careful selection and integration of different tools in immune cell analysis for a more complete understanding of complex tumor immunology.

References

[1] Zhang Z, Bao S, Yan C, Hou P, Zhou M, Sun J. Computational Principles and Practice for Decoding Immune Contexture in the Tumor Microenvironment. Brief Bioinform, 2020.

[2] Zeng D, Ye Z, Shen R, et al. IOBR: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures [J]. Frontiers in immunology, 2021.

[3] Ma W, Wei S, Long S, Tian EC et al. Dynamic evaluation of blood immune cells predictive of response to immune checkpoint inhibitors in NSCLC by multicolor spectrum flow cytometry. Front Immunol 2023.

[4] Li B, Severson E, Pignon J C, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy[J]. Genome biology, 2016.

[5] Finotello F, Mayer C, Plattner C, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data[J]. Genome medicine, 2019.

[6] Newman A M, Liu C L, Green M R, et al. Robust enumeration of cell subsets from tissue expression profiles [J]. Nature methods, 2015.

[7] Lv H, Liu X, Zeng X, et al. Comprehensive analysis of cuproptosis-related genes in immune infiltration and prognosis in melanoma[J]. Frontiers in pharmacology, 2022.

[8] Jiang Y, Chen Z, Han N, et al. sc-ImmuCC: hierarchical annotation for immune cell types in single-cell RNA-seq[J]. Frontiers in Immunology, 2023.

[9] Sun J, Zhang Z, Bao S, et al. Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer [J]. Journal for immunotherapy of cancer, 2020.

[10] Kim S I, Cassella C R, Byrne K T. Tumor burden and immunotherapy: impact on immune infiltration and therapeutic outcomes [J]. Frontiers in immunology, 2021.