基于图神经网络的肺癌单细胞转录组数据细胞类型识别*

沈　阳[1]　王　彤[1,2]　刘改琴[1]　田雅昕[1,2]　房瑞玲[1,3]　罗艳虹[1,3]　曾　平[4]
余红梅[1,3]　张岩波[1,3]　曹红艳[1,3△]

【提　要】 **目的**　探讨单细胞图神经网络（single-cell graph neural network，scGNN）方法在肺癌单细胞转录组测序（scRNA-seq）数据上的应用，识别肺癌细胞亚群，发现肺癌潜在靶点及特异性标志物。**方法**　采用 scGNN 方法对肺癌 scRNA-seq 数据进行聚类分析，并与经典的单细胞聚类方法 SC3、Seurat 以及谱聚类方法进行比较。进一步对 scGNN 方法的细胞分型结果进行细胞注释、差异基因筛选、GO 富集分析和细胞通讯等下游分析。**结果**　scGNN 将肺癌 scRNA-seq 数据细胞分为 10 类，并将其注释为五种主要细胞类型。筛选出 841 个差异基因，GO 富集分析得到 715 个富集项。细胞通讯分析检测到 118 个重要的配体-受体对，14 条信号通路，并分析 MIF 信号通路网络中细胞类型间的相互作用。**结论**　scGNN 方法能够有效地识别细胞亚型，其细胞分型准确性优于传统分型方法，通过筛选差异表达基因，探索细胞间相互作用，识别潜在生物标志物，获取肺癌相关靶点，对肺癌靶向药物治疗策略具有重要意义。

【关键词】 单细胞转录组　图神经网络　肺癌　细胞类型识别
【中图分类号】 Q78；R734.2　　【文献标志码】 A　　**DOI**　10.11783/j.issn.1002-3674.2026.05.006

Cell Type Identification of Single-cell RNA Sequencing Data in Lung Cancer based on Graph Neural Network

Shen Yang, Wang Tong, Liu Gaiqin ,et al(*Division of Health Statistics*, *Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment*, *Shanxi Medical University*, *Taiyuan* 030001)

【**Abstract**】 **Objective**　This study explores the application of graph neural networks (GNNs) on lung cancer scRNA-seq data to identify potential therapeutic targets and specific biomarkers, providing valuable insights for targeted therapy. **Methods**　The scGNN method is used to analyze the data and is compared with three other approaches. Further downstream analyses, including cell annotation, differential gene screening, GO enrichment analysis, and cell-cell communication, are also performed. **Results**　The scGNN method divides the sample data into 10 clusters and identifies them as five major cell types. A total of 841 differential expression genes are selected, with the top 15 visualized. GO enrichment analysis yielded 715 significant terms, with the top 15 most enriched terms presented graphically. Cell-cell communication analysis detects 118 important ligand-receptor pairs, which are further categorized into 14 signaling pathways. The MIF signaling pathway is analyzed in detail. **Conclusion**　The scGNN method effectively identifies cell subtypes and outperforms traditional classification methods in terms of results. Subsequent analysis can help identify potential biomarkers and lung cancer-related targets, which is of great significance for tumor-targeted drug treatment strategies.

【**Key words**】 Single-cell transcriptomics; Graph neural networks; Lung cancer; Cell type identification

肺癌是最常见的恶性肿瘤，也是全球第一大癌症，发病率和死亡率呈上升趋势[1-2]。相关研究[3]指出，2024 年中国新发肺癌病例约 117.59 万例，死亡病例约 74.33 万例。肺癌是一种通过多种途径发展的异质性疾病，涉及多种基因、通路及免疫等分子机制的改变，其高度异质性是影响靶向药物疗效的重要原因，也是治疗难以取得突破性进展的主要因素[4-5]。传统的批量测序方法关注所有细胞的平均表达水平，难以揭示肺癌中不同细胞群体的异质性特征，而单细胞转录组测序（single cell RNA sequencing，scRNA-seq）技术检测单个细胞的基因表达水平，将分辨率由传统多个细胞的平均水平提升到单个细胞水平，能捕捉到肺癌组织细胞水平特异性变化，为肺癌临床诊疗提供了新的方向[6]。

以 scRNA-seq 数据为基础分析细胞异质性[7]，识别细胞亚群，能有效揭示细胞间异质性和功能多样性。scRNA-seq 数据具有高维度、高稀疏、高噪声等特点，细胞类型识别在统计建模和计算上面临挑战[8]。现有的细胞类型识别方法如谱聚类[9]通过构建相似性矩阵并计算其特征向量进行聚类。SC3[10]通过计算欧氏距离、Pearson 和 Spearman 相关系数构建距离矩阵，应用主成分分析（principal component analysis，PCA）降维后对特征向量进行 k-means 聚类，然后，利用 CSPA（cluster-based similarity partitioning algorithm）算法计算共识矩阵，最后用层次聚类对共识矩阵进行

* 基金项目：国家自然科学基金（82473739；82273742）；山西省基础研究计划（202303021211130）；山西省回国留学人员科研资助项目（2024-081）；山西省高等教育“百亿工程”科技引导专项；煤炭环境致病与防治教育部重点实验室（山西医科大学）开放课题基金（MEKLCEPP/SXMU-202415）
1. 山西医科大学卫生统计教研室 重大疾病风险评估山西省重点实验室（030001）
2. 山西医科大学医学科学院
3. 煤炭环境致病与防治教育部重点实验室
4. 徐州医科大学公共卫生学院流行病与卫生统计学系
△ 通信作者：曹红艳，E-mail：caohy@sxmu.edu.cn

聚类得到最终的细胞分型结果。Seurat[11]通过标化、降维（如PCA）及高变基因的选择，将高维基因表达数据映射到低维空间，进而应用图聚类算法对细胞进行聚类。这些方法通过抽样策略降低了细胞数量，可能会降低细胞类型识别的准确性，同时也存在忽略罕见细胞类型的风险[12]。自编码器作为深度学习中的无监督方法，通过对数据编码重构，将其映射到低维空间，能有效地捕获特征间的非线性关系。基于图神经网络的单细胞聚类方法scGNN（single-cell graph neural network）[13]能够在全局视角下学习图拓扑的低维表示和训练节点之间的关系，其可用于揭示细胞间异质性及潜在的复杂基因表达模式。通过结合基因表达和转录调控信息，scGNN基于拓扑提取（topological abstraction）训练低维特征向量（即嵌入），从而表示细胞之间的关系。该模型通过综合细胞群体的基因表达和细胞间相互作用信息，提供一个全局视角，提高了细胞类型识别的准确性和稳健性。

因此，本文采用scGNN方法，对肺癌scRNA-seq数据进行细胞类型识别，并将其与SC3、Seurat以及谱聚类方法进行比较，验证其在细胞类型识别中的优越性。基于细胞聚类结果，进一步开展下游分析，挖掘关键基因或信号通路，为肺癌患者靶向治疗方案制定提供理论依据。

数据和方法

1. 数据来源与处理

从高通量基因表达（gene expression omnibus，GEO）数据库[14]（https://www.ncbi.nlm.nih.gov/geo/）下载GSE164146的数据，该数据包含4例非小细胞肺癌（non-small cell lung cancer，NSCLC）患者的肿瘤与癌旁组织样本，本研究使用了4个肿瘤样本，共4887个细胞。利用R软件中的Seurat包分别将四例样本数据矩阵转化为seurat对象[15]，保留至少在3个细胞中表达的基因与至少有200个基因表达的细胞，对四例样本进行合并与质量控制，保留线粒体基因含量<15%，总基因>20000个，基因表达量为600～4000的细胞，最终得到的肺癌细胞样本包含3933个细胞，10138个基因，并对数据进行归一化处理。

2. scGNN方法

基于图神经网络（graph neural network，GNN）的图自编码器，其独特优势在于能够学习图拓扑的低维表示，并从全局视角训练节点之间的关系[16]。scGNN方法不为基因表达数据或“dropout”事件假定任何统计分布或关系，直接使用自底向上的方式聚合细胞间关系，在多个自编码器的基础上，通过添加正则化项以整合基因调控信号（gene regulatory signals，TRSs），反复迭代获得最终的细胞聚类结果（图1）。scGNN接收scRNA-seq表达矩阵作为输入，输出细胞聚类结果与细胞网络。算法迭代过程为：结合表达矩阵和基因调控信号提取低维表征，构建K近邻（K-nearest neighbor，KNN）图；随后利用图自编码器学习图嵌入开展细胞聚类。一旦聚类结果收敛或达到最大迭代次数，迭代终止。如未达到收敛条件，则根据当前聚类结果更新特征自编码器得到重构表达，重新构造KNN图，再次运行图自编码器并更新聚类，反复迭代直至收敛，最终输出聚类结果和细胞网络。

（1）特征自编码器与正则化

特征自编码器，即在编码器和解码器中堆叠两层全连接网络，旨在学习scRNA-seq数据的主要特征表示。对于输入的基因表达数据X，编码器首先将其映射为低维嵌入X'，然后再将X'重构为$\hat{X}$，$X, \hat{X} \in \mathbb{R}^{N \times M}$，$X' \in \mathbb{R}^{N \times M'}$。其中$M$为输入基因的数量，$M'$为学习到的嵌入维度，$M' < M$。训练特征自编码器的目标是通过最小化损失函数，最大化原始表达与重构表达之间的相似

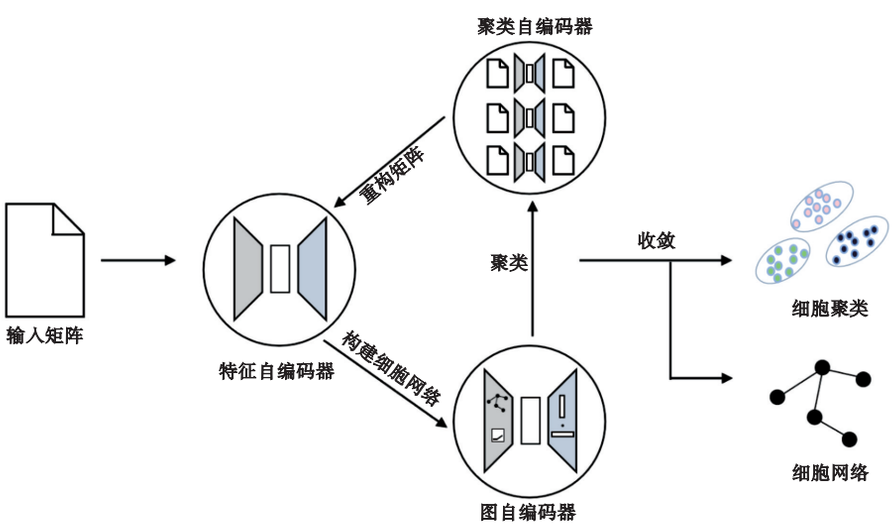


图1 scGNN方法原理示意图

性。均方误差(mean squared error, MSE)定义如下：

$$\alpha\sum\left(\left(X-\hat{X}\right)^2\circ \mathrm{TRS}\right) \tag{1}$$

其中，$\mathrm{TRS}\in\mathbb{R}^{N\times M}$，$\alpha\in[0,\ 1]$是控制基因调控正则化强度的参数，$\circ$表示逐元素相乘。因此，特征自编码器的损失函数如式(2)所示：

$$Loss=(1-\alpha)\sum\left(X-\hat{X}\right)^2+\alpha\sum\left(\left(X-\hat{X}\right)^2\circ \mathrm{TRS}\right) \tag{2}$$

在编码器部分，第一层和第二层的输出维度分别设置为 512 和 128，且每一层均使用 ReLU 激活函数，$\mathrm{ReLU}(x)=\max(0,\ x)$。在解码器部分，第一层和第二层的输出维度分别为 128 和 512，每一层使用 sigmoid 激活函数，$\mathrm{sigmoid}(x)=1/(1+e^{-x})$。学习率设定为 0. 001。

(2)细胞网络和剪枝

细胞网络利用从特征自编码器中学习到的嵌入表示细胞间关系。使用 KNN 方法构建细胞图(使用欧氏距离构建图，默认 K=10)，其中节点代表单个细胞，边则表示细胞之间的相互关系[17]。使用孤立森林[18]对网络进行修剪，通过构建一组独立的随机森林模型，对节点与其 K 个最近邻之间的距离关系进行评估分析。算法采用动态阈值策略，自动估计数据中的异常比例，将识别出的异常节点连接予以断开。

(3)图自编码器

图自编码器的编码器以修剪后的细胞网络为输入，使用图卷积网络(graph convolution network, GCN)学习每个节点的低维嵌入，然后通过图自编码器的解码器重新生成整个图结构。使用 k-means 聚类方法对学习到的图嵌入上的细胞进行聚类，簇的数量由输入细胞网络上的 Louvain 算法确定[19]，本研究中默认基于网络特性自动选择 Louvain 算法分辨率。

Louvain 算法是一种基于模块度优化的社区发现方法，结合了贪婪策略的高效性和模块度优化的准确性。算法分为两个阶段迭代执行：首先，每个节点遍历邻居社区，选择模块度增益最大的社区进行合并，直到无法提升；然后将社区合并为超级节点，边权重为原社区间连接之和。重复这一过程直到模块度不再增加，最终得到层次化的社区结构。模块度 Q 与增量 ΔQ 公式如下：

$$Q=\sum_c\left[\frac{\sum_{in}}{2m}-\left(\frac{\sum_{tot}}{2m}\right)^2\right] \tag{3}$$

$$\Delta Q=\left[\frac{\sum_{in}+k_{i,in}}{2m}-\left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right]-\left[\frac{\sum_{in}}{2m}-\left(\frac{\sum_{tot}}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right] \tag{4}$$

其中 $\sum_{in}$ 表示社区 c 内部的权重，$\sum_{tot}$ 表示的是与社区 c 内部的点连接的边的权重，m 是网络中所有链路的权重总和，$k_{i,in}$ 为所有从节点 i 指向社区 c 的边的权值和，k_i 为指向节点 i 的所有边的权值和。

设输入的修剪后细胞网络为 $G=(V,\ E)$，其中节点集 $N=|V|$ 表示细胞，边集 E 表示细胞间关系。A 表示该网络的邻接矩阵，D 为度矩阵。GCN 定义为：$\mathrm{GCN}(X',\ A)=\mathrm{ReLU}(\tilde{A}XW)$，其中 W 是从训练中学习到的权重矩阵，$\tilde{A}=D^{-1/2}AD^{-1/2}$ 为对称归一化邻接矩阵。

图自编码器的编码器部分由两层 GCN 组成，在式(3)中，Z 是编码器所学到的图嵌入，W_1 和 W_2 分别为第一层和第二层的权重矩阵，这两层的输出维度分别设置为 32 和 16，学习率设为 0. 001。

$$Z=\mathrm{ReLU}\left(\tilde{A}ReLU(\tilde{A}XW_1)W_2\right) \tag{5}$$

图自编码器的解码器部分定义为嵌入之间的内积，如式(6)：

$$\hat{A}=\mathrm{sigmoid}(ZZ^T) \tag{6}$$

其中，$\hat{A}$ 是 A 的重构邻接矩阵。

学习图自编码器的目标是最小化输入邻接矩阵 A 和重构矩阵 $\hat{A}$ 之间的交叉熵 L：

$$L\left(A,\hat{A}\right)=-\frac{1}{N\times N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(a_{ij}\log(\hat{a}_{ij})+(1-a_{ij})\log(1-\hat{a}_{ij})\right) \tag{7}$$

其中，a_{ij} 和 $\hat{a}_{ij}$ 是邻接矩阵 A 和 $\hat{A}$ 中第 i 行第 j 列的元素。由于网络中共有 N 个细胞节点，因此 $N\times N$ 为邻接矩阵中元素的总数。

(4)聚类自编码器

通过聚类自编码器，重建每个细胞聚类中来自特征自编码器的表达矩阵。利用图自编码器推断的细胞类型信息，每一细胞簇中的细胞通过单独的聚类自编码器来重建基因表达值。通过迭代，重建的矩阵被反馈到特征自编码器中，直到收敛。

(5)迭代过程

细胞网络的迭代过程可以定义为：

$$\tilde{A}=\lambda L_0+(1-\lambda)\frac{A_{ij}}{\sum_j A_{ij}} \tag{8}$$

其中，L_0 为初始剪枝网络的归一化邻接矩阵，$L_0={D_0}^{-1/2}A_0{D_0}^{-1/2}$，$D_0$ 为度矩阵；$\lambda\in[0,\ 1]$为控制收敛速度的参数。在 t 次迭代中，可通过判断邻接矩阵是否收敛，即 $\tilde{A}_t-\tilde{A}_{t-1}<\gamma\tilde{A}_0$，来决定是否停止迭代。设置 $\lambda=0.5$，$\gamma=0.99$。将最后一次迭代得到的聚类结果作为最终的细胞类型结果。

3. 方法比较

使用 scGNN、SC3、Seurat 和谱聚类方法对预处理

后的数据进行聚类分析，并比较它们的聚类性能。其中，scGNN以基因表达矩阵为输入，选取方差最大的2000个基因进行后续研究，并对输入数据进行对数转换。使用*t*-分布随机邻域嵌入（*t*-distributed stochastic neighbor embedding，*t*-SNE）[20]可视化、Davies-Bouldin指数（Davies-Bouldin index，DBI）[21]及Calinski-Harabasz（CH）[22]指数对聚类结果的质量进行评价。

DBI计算公式为：

$$\mathrm{DBI}=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(C_i+C_j)}\right) \tag{9}$$

其中，k为聚类数，σ_i为第i个簇内所有样本到该簇中心的平均距离，$d(C_i+C_j)$表示第i簇和第j簇之间的距离，通常用簇中心之间的欧氏距离衡量。DBI值越小表示聚类质量越好。

CH计算公式为：

$$\mathrm{CH}=\frac{tr(B_k)(n-k)}{tr(W_k)(k-1)} \tag{10}$$

其中，k为聚类数，n为数据点总数，B_k为簇间的协方差矩阵，W_k为簇内的协方差矩阵。CH值越大表示聚类质量越好。

4. 细胞注释

采用R软件中的SingleR包[23]进行细胞注释，使用软件包中默认的Spearman相关性方法，基于聚类结果和log归一化表达矩阵进行注释，未显式设置相关性阈值，但通过“quantile”参数（默认0.8）过滤低置信注释。参考基因列表为人类原代细胞图谱（human primary cell atlas，HPCA）数据集[24]，共713例样本，涵盖36种主要细胞类型，对细胞进行主要类别的细胞注释。

5. 差异基因筛选与GO富集分析

差异表达基因（differentially expressed genes，DEGs）的筛选使用R软件中的Seurat包[11]进行分析。具体而言，使用“FindAllMarkers”函数通过Wilcoxon秩和检验比较不同细胞群体间的基因表达差异，筛选标准设定为：$P_{adj}<0.001$、$|\log_2(foldchange)|>2$，用以筛选差异表达的基因。基因本体（gene ontology，GO）富集分析[25]则是使用clusterProfiler包[26]，依赖org. Hs. eg. db数据库[27]完成的，将阈值设置为pvalueCutoff=0.5，qvalueCutoff=0.5，系统分析差异基因在细胞组分（cellular component，CC）、分子功能（molecular function，MF）和生物学过程（biological process，BP）三个层面的功能特征。

6. 细胞通讯

为了研究细胞间的通讯，使用R软件中的CellChat包[28]识别并量化不同细胞类型之间的相互作用，通过将细胞的基因表达数据和细胞类型信息作为输入，进行配体-受体分析。所使用的配体-受体库来源于CellChatDB. human数据集。通过可视化图展示了不同细胞类型之间的通讯强度和模式。

结　果

1. 四种方法聚类结果比较

采用scGNN、SC3、Seurat和谱聚类方法对基因表达矩阵进行聚类分析。四种方法得到的聚类结果如图2所示。

图2为四种聚类方法的*t*-SNE可视化结果。scGNN方法通过Louvain算法在细胞图上自动确定簇的数量，最终将细胞分为10类。SC3方法将细胞分为10类，Seurat方法将细胞分为18类，谱聚类将细胞分为10类。表1列出了四种方法聚类结果的DBI和CH指数，结果显示，scGNN方法的DBI值最低、CH值最高，表明其聚类效果优于SC3、Seurat及谱聚类方法。

2. 细胞类型识别

依据scGNN聚类结果中每个细胞亚群的特异性高表达基因，进一步对聚类结果进行细胞注释。最终识别出5种主要细胞亚群，包括内皮细胞（endothelial cells）、上皮细胞（epithelial cells）、巨噬细胞（macrophages）、单核细胞（monocytes）和T细胞（T cells），如图3A所示，其中T细胞进一步细分为初始CD8+T细胞（T_cell:CD8+_naïve）。结合图2A可知，簇2、4、5和7主要包含上皮细胞，簇9主要由内皮细胞构成，簇0和簇8则主要为巨噬细胞，簇1和簇6主要为单核细胞，簇3则以T细胞为主。图3B展示了总细胞中五种类型的相对比例，其中上皮细胞占比最高，具体细胞数量及占比详见表2。图3C则展示了四例肿瘤细胞中五种类型的数量及比例。

3. 差异基因筛选

为揭示不同细胞类型之间的差异基因，共筛选出841个DEGs，并计算每个差异基因在各组中的平均表达值，结果通过可视化展示。图4A展示了前15个差异基因及其在相应细胞类型中的表达情况，分析结果表明，*EPCAM*、*CD*24、*PSCA*、*CDH*1及*COL*1*A*1为上皮细胞的高表达基因，*FCN*1、*S*100*B*、*VCAN*为单核细胞的高表达基因，*CD*8*A*、*CCR*7、*CD*79*A*及*JCHAIN*为T细胞的高表达基因，*FCGR*1*A*基因在巨噬细胞内高表达，*TCF*4和*PECAM*1基因在内皮细胞中高表达。图4B随机选取了部分差异基因，比较了它们在不同细胞类型中的表达差异。将前15个差异基因进行GO富集分析共得到715个富集项，最显著的前15个富集项如图4C所示，点大小表示该项目的基因数量，点颜色表示该项的调整*P*值，主要涉及细胞识别（GO：0008037：cell recognition）、免疫应答调节的信号通路（GO：0002764：immune response-regulating signaling pathway）等。

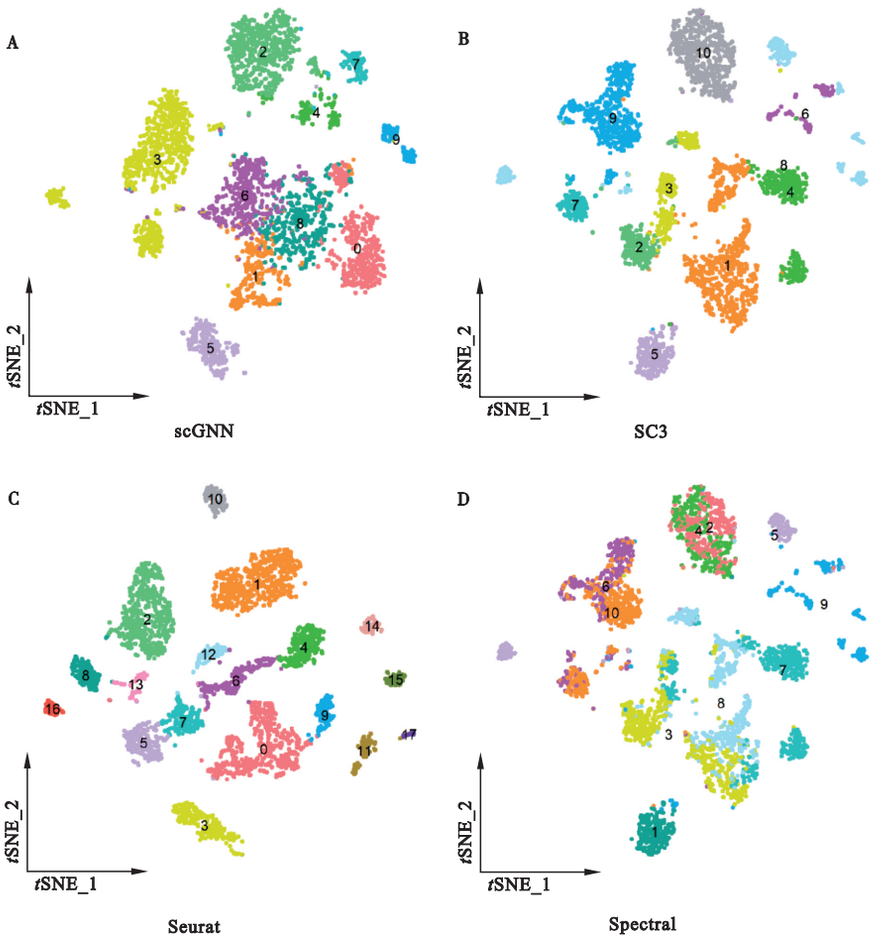


图 2　四种聚类方法的结果可视化图

表 1　四种聚类方法评价指标值

细胞聚类方法	聚类数	DBI	CH
scGNN	10	4. 30	103. 00
SC3	10	5. 41	97. 37
Seurat	18	5. 09	60. 38
Spectral	10	6. 14	98. 11

表 2　4 例肿瘤细胞样本中不同细胞类型数量及比例

细胞类型	细胞数量	细胞比例(%)
内皮细胞	126	3. 20
上皮细胞	1180	30. 00
巨噬细胞	918	23. 34
单核细胞	825	20. 98
T 细胞	884	22. 48

4. 细胞通讯分析

图 5A 和 5B 分别为不同细胞类型间相互作用的数量和权重/强度。其中，节点代表不同的细胞类型，其大小与细胞数量成正比，图中显示上皮细胞数量最多，而内皮细胞数量最少。箭头指示作用的方向。边的粗细代表数量/权重，尤以上皮细胞与巨噬细胞之间的相互作用最为强烈。自环则表示某一细胞类型与自身相互作用的数量或强弱。

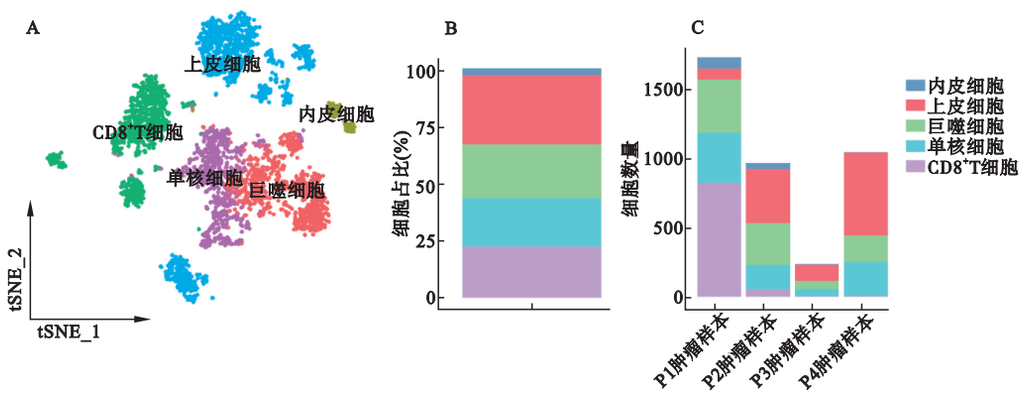


图 3　NSCLC 细胞类型识别

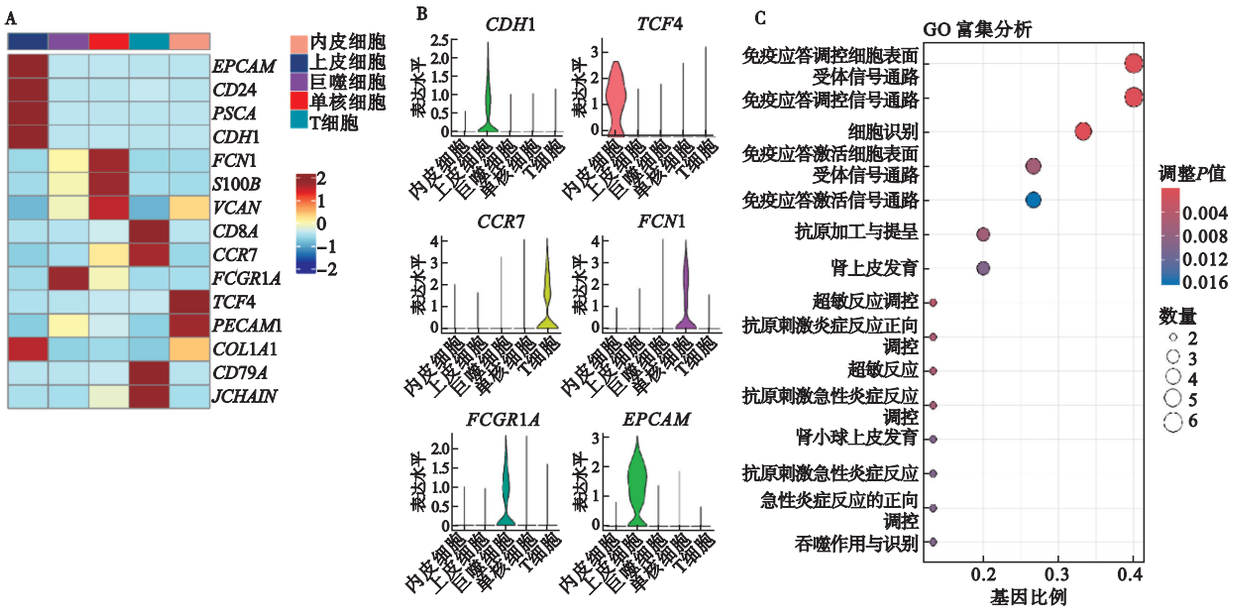


图4　差异表达基因和GO富集分析

在五种细胞类型中共鉴定到118个重要的配体–受体对，这些配体–受体对进一步划分为14条信号通路，包括*MIF*、*CypA*、*GALECTIN*、*MK*、*VISFATIN*、*ANNEXIN*、*RESISTIN*、*EGF*、*SPP*1、*UGRP*1、*GDF*、*TWEAK*、*TGFb*和*TRAIL*。分层图5C为MIF信号通路网络，左侧为巨噬细胞、单核细胞和上皮细胞的自分泌和旁分泌信号，右侧则为其他类型细胞的信号。实心圆和空心圆分别代表信号源和目标，圆圈大小与每类细胞的数量成正比。边的宽度反映通讯概率，边的颜色与信号源一致。通过MIF通路的环形图（图5D）可知MIF信号通路在上皮细胞与单核细胞的通讯中起着核心作用。图5E为各配体–受体对在MIF信号通路整体通信网络中的相对贡献，MIF的致癌作用是通过*CD*74、*CD*44和*CXCR*4结合介导的。图5F则为基于MIF信号通路网络中的四个度量指标（发送细胞、接收细胞、介导细胞、调控细胞）计算的细胞类型相对重要性热图。

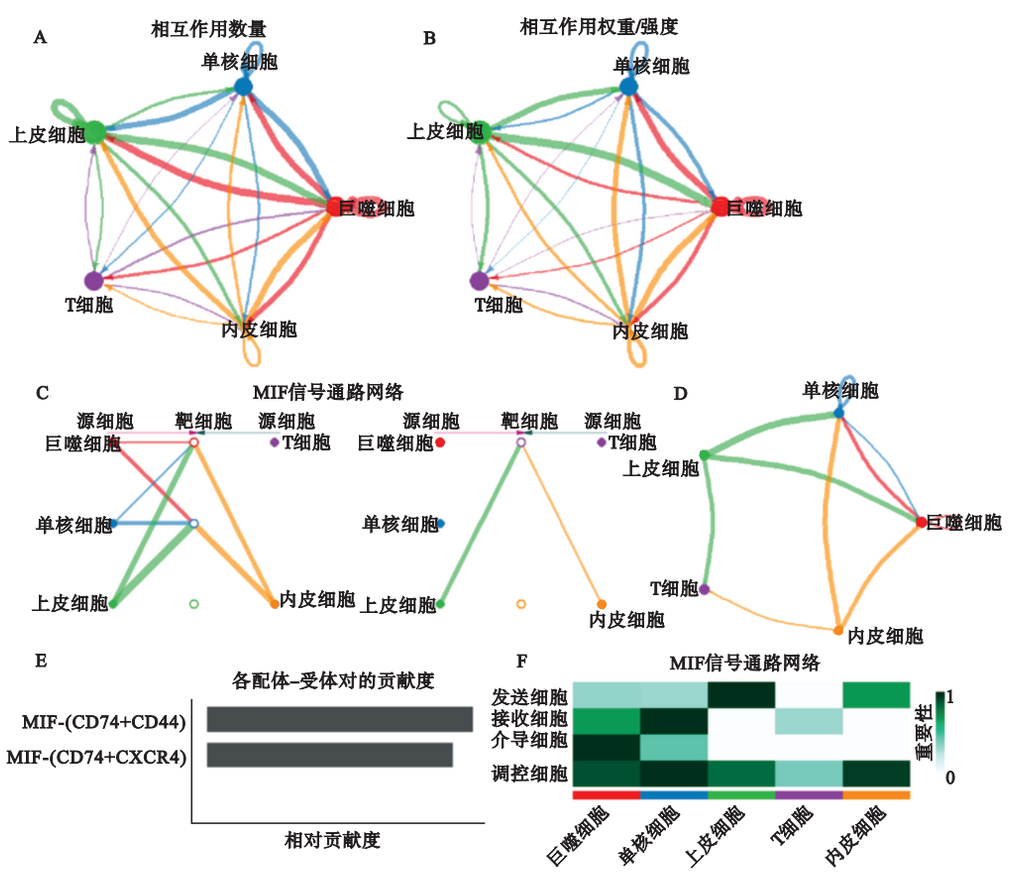


图5　细胞通讯与MIF信号网络可视化

讨 论

本文采用 scGNN 方法对肺癌单细胞转录组学数据进行细胞类型识别，并将其与另外三种单细胞聚类方法进行比较，再通过细胞注释，将细胞分为五种不同类型。基于细胞亚型得到的潜在生物标记物，如特定基因或信号通路，将为肺癌靶向治疗提供参考依据。

scGNN 对肺癌的细胞识别结果相较于 SC3、Seurat 和谱聚类方法更为准确。其中，scGNN 将细胞聚类为 10 类，细胞注释结果显示五种主要类型细胞：内皮细胞、上皮细胞、巨噬细胞、单核细胞和 T 细胞，这一高分辨率的分类得益于 scGNN 的图神经网络架构，但细胞注释的准确性也依赖于参考数据集的质量。例如，本研究使用的 HPCA 数据集覆盖广泛的正常原代细胞类型，其标准化 logcounts 数据适合跨平台比对，为 scGNN 提供了稳定的注释基线[24]。然而，HPCA 在肺癌单细胞数据分析中可能存在缺乏肿瘤微环境特异性标签、细胞亚群分辨率有限、批次效应影响等问题[29]。本研究细胞注释结果中上皮细胞占比最高，实际约 90% 的肺癌病例与吸烟相关，吸烟引起的遗传损伤可促使上皮细胞癌变，进而导致肺癌发生[30]。未来研究可结合肺癌特异性参考集提高肿瘤微环境注释准确性，或进行多数据库整合以减少偏差。

此外，在差异基因筛选分析中，*EPCAM*、*CD*24、*FCGR*1*A*、*CCR*7 及 *TCF*4 等基因的表达均显著上调。其中，*EPCAM* 是一种细胞黏附分子，能够介导上皮特异性、钙离子独立的同源细胞间黏附，且作为首个被发现的人类肿瘤相关抗原，目前已被广泛应用于癌症免疫治疗靶点的研究[31]。*CD*24 具有强效的抗吞噬功能，能够直接保护癌细胞免受表达 *Siglec*−10 的巨噬细胞的攻击，单克隆抗体阻断 *CD*24 − *Siglec*−10 信号可强力增强对 *CD*24+ 肿瘤的清除，这表明 *CD*24 阻断在免疫疗法中大有可为[32]。以往研究显示，在多种癌症中，*FCGR*1*A* 的表达水平与更好的预后相关[33]。因此，*FCGR*1*A* 被认为是一种潜在的预后生物标志物。*CCR*7 在促进淋巴系统转移、协调免疫反应中发挥着重要作用[34]，被证明与更好的预后相关，过表达的患者生存结果会更好[35]。*TCF*4 作为癌基因，能够促进癌症迁移、侵袭以及增加耐药性，并在多种癌症中上调，因此被视为预示不良预后的重要风险因子[36]。

本研究的 GO 富集分析显示，肺癌微环境中 cell recognition（GO：0008037）和 immune response−regulating signaling pathway（GO：0002764）显著激活，提示肿瘤细胞可能通过免疫识别逃逸机制（如 MHC 下调或 PD−L1 上调）规避 T 细胞杀伤[37]。同时，免疫调控通路的激活（如 JAK−STAT 或 NF−κB）可能驱动 TAMs 向 M2 型极化，并通过 SPP1−CD44 互做促进免疫抑制性微环境形成[38]。

最后，本研究基于细胞通讯分析，着重探讨了 MIF 信号通路网络在肺癌中的作用。在肿瘤中，巨噬细胞通常分化为促进肿瘤发展的 M2 型肿瘤相关巨噬细胞（TAM），从而加速肿瘤细胞的增殖和转移，并抑制抗肿瘤免疫应答[39−40]，而 MIF 可通过调节 TAM 的极性，抑制肿瘤增殖和转移。MIF 还是肿瘤微环境中的一种免疫调节剂，能诱导肿瘤抑制性的免疫微环境形成[39]。其他通路也在肿瘤微环境中表现出复杂的调控网络，例如 CypA−CD147 轴通过激活 NF−κB 信号促进肿瘤侵袭和转移[41]，而 GALECTIN−3 通过调控 T 细胞耗竭和 Treg 功能影响免疫治疗响应[42]，EGF 信号通过 RAS−MAPK 和 PI3K−AKT 通路驱动肿瘤增殖[43]，此外，TGFβ 的双重作用[44]和 SPP1+ 巨噬细胞的促瘤机制[45]进一步强调了靶向细胞间通讯在癌症治疗中的潜力。

综上所述，本文基于 scGNN 方法对肺癌 scRNA−seq 数据进行细胞类型识别，结果显示该方法在亚型识别方面优于三种以往方法，为单细胞测序数据分析提供了新的视角。通过筛选差异表达基因，识别肺癌相关靶点，并结合细胞类型识别与细胞通讯分析，进一步探索了细胞间相互作用，揭示了关键通路并识别了潜在生物标志物，这些发现为肿瘤靶向药物的治疗策略提供了重要理论依据。

参 考 文 献

[1] Smolarz B, Łukasiewicz H, Samulak D, et al. Lung cancer: epidemiology, pathogenesis, treatment and molecular aspect (review of literature)[J]. International journal of molecular sciences, 2025, 26(5): 2049.

[2] Thapa R, Gupta G, Bhat AA, et al. A review of Glycogen Synthase Kinase-3 (GSK3) inhibitors for cancers therapies[J]. International Journal of Biological Macromolecules, 2023, 253(Pt7): 127375.

[3] 孙可欣,李荔,王少明,等. 2024年中国分地区恶性肿瘤流行情况分析[J]. 中华肿瘤杂志,2026,48(3):400-412.

[4] Bai X, Li Y, Zeng X, et al. Single-cell sequencing technology in tumor research[J]. Clinica Chimica Acta, 2021, 518: 101-109.

[5] 张凌谱,杨佳臻,唐东升,等. 靶向肿瘤进化的创新纳米药物[J]. 中国科学:化学,2022,52(12):2121-2155.

[6] 刘亚飞, 叶贯超, 黄岚, 等. 单细胞转录组在晚期肺腺癌患者个体化治疗中的应用[J]. 中华实验外科杂志, 2020, 37(5): 3.

[7] Wang Z, Ding H, Zou Q. Identifying cell types to interpret scRNA-seq data: how, why and more possibilities[J]. Briefings in Functional Genomics, 2020, 19(4): 286-291.

[8] Hu J, Li X, Hu G, et al. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis[J]. Nature Machine Intelligence, 2020, 2(10): 607-618.

[9] Park S, Zhao H. Spectral clustering based on learning similarity

matrix[J]. Bioinformatics, 2018, 34(12): 2069-2076.

[10] Kiselev VY, Kirschner K, Schaub MT, et al. SC3: consensus clustering of single-cell RNA-seq data[J]. Nature Methods, 2017, 14(5): 483-486.

[11] Satija R, Farrell JA, Gennert D, et al. Spatial reconstruction of single-cell gene expression data[J]. Nature Biotechnology, 2015, 33 (5): 495-502.

[12] 何佳, 王文杰, 张刘超, 等. Specter聚类方法在单细胞RNA测序数据分析中的应用[J]. 中国卫生统计, 2023, 40(3): 354-357.

[13] Wang J, Ma A, Chang Y, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses[J]. Nature Communications, 2021, 12(1): 1882.

[14] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets: update[J]. Nucleic Acids Research, 2012, 41(D1): D991-D995.

[15] Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis[J]. Nature Biotechnology, 2024, 42(2): 293-304.

[16] Kipf TN, Welling M. Variational Graph Auto-Encoders[J]. Stat, 2016, 1050: 21.

[17] Wolf FA, Hamey FK, Plass M, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells[J]. Genome Biology, 2019, 20(1): 59.

[18] Liu FT, Ting KM, Zhou ZH. Isolation forest[C]//2008 eighth ieee international conference on data mining. IEEE, 2008: 413-422.

[19] Blondel V, Guillaume JL, Lambiotte R. Fast unfolding of communities in large networks: 15 years later[J]. Journal of Statistical Mechanics: Theory and Experiment, 2024, 2024(10): 10R001.

[20] Linderman GC, Rachh M, Hoskins JG, et al. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data [J]. Nature Methods, 2019, 16(3): 243-245.

[21] Davies DL, Bouldin DW. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979 (2): 224-227.

[22] Caliński T, Harabasz J. A dendrite method for cluster analysis[J]. Communications in Statistics-theory and Methods, 1974, 3(1): 1-27.

[23] Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage [J]. Nature Immunology, 2019, 20(2): 163-172.

[24] Mabbott NA, Baillie JK, Brown H, et al. An expression atlas of human primary cells: inference of gene function from coexpression networks[J]. BMC Genomics, 2013, 14(1): 632.

[25] Gene Qntology Consortium. The Gene Ontology resource: enriching a GOld mine[J]. Nucleic Acids Research, 2021, 49(D1): D325-D334.

[26] Wu T, Hu E, Xu S, et al. clusterProfiler 4. 0: A universal enrichment tool for interpreting omics data[J]. The Innovation, 2021, 2 (3): 100141.

[27] Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor[J]. Nature Methods, 2015, 12(2): 115-121.

[28] Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell-cell communication using CellChat[J]. Nature Communications, 2021, 12(1): 1088.

[29] Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities[J]. Experimental & Molecular Medicine, 2020, 52(9): 1452-1465.

[30] Rooney C, Sethi T. The epithelial cell and lung cancer: the link between chronic obstructive pulmonary disease and lung cancer[J]. Respiration, 2011, 81(2): 89-104.

[31] Pak MG, Shin DH, Lee CH, et al. Significance of EpCAM and TROP2 expression in non-small cell lung cancer[J]. World Journal of Surgical Oncology, 2012, 10: 1-8.

[32] Barkal AA, Brewer RE, Markovic M, et al. CD24 signalling through macrophage Siglec-10 is a target for cancer immunotherapy [J]. Nature, 2019, 572(7769): 392-396.

[33] Xu J, Guo Y. FCGR1A serves as a novel biomarker and correlates with immune infiltration in four cancer types[J]. Frontiers in Molecular Biosciences, 2020, 7: 581615.

[34] Salem A, Alotaibi M, Mroueh R, et al. CCR7 as a therapeutic target in Cancer[J]. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2021, 1875(1): 188499.

[35] Yue Z, Ningning D, Lin Y, et al. Correlation between CXCR4, CXCR5 and CCR7 expression and survival outcomes in patients with clinical T1N0M0 non-small cell lung cancer[J]. Thoracic Cancer, 2020, 11(10): 2955-2965.

[36] Wang H, Sahu A, Chuong MD, et al. TCF4 as a potential prognostic biomarker and an anticancer target in gastric cancer[J]. Translational Cancer Research, 2024, 13(9): 5073.

[37] He Y, Yu F, Tian Y, et al. Single-cell RNA sequencing unravels distinct tumor microenvironment of different components of lung adenocarcinoma featured as mixed ground-glass opacity[J]. Frontiers in Immunology, 2022, 13: 903513.

[38] Deng Y, Xia L, Zhang J, et al. Multicellular ecotypes shape progression of lung adenocarcinoma from ground-glass opacity toward advanced stages[J]. Cell Reports Medicine, 2024, 5(4): 101489.

[39] Calandra T, Bucala R. Macrophage migration inhibitory factor (MIF): a glucocorticoid counter-regulator within the immune system [J]. Critical Reviews™ in Immunology, 1997, 17(1): 77-88.

[40] Tilstam PV, Schulte W, Holowka T, et al. MIF but not MIF-2 recruits inflammatory macrophages in an experimental polymicrobial sepsis model[J]. The Journal of Clinical Investigation, 2021, 131(23): e127171.

[41] Han JM, Jung HJ. Cyclophilin A/CD147 interaction: A promising target for anticancer therapy[J]. International Journal of Molecular Sciences, 2022, 23(16): 9341.

[42] Chen Y, Wang D, Li Y, et al. Spatiotemporal single-cell analysis decodes cellular dynamics underlying different responses to immunotherapy in colorectal cancer[J]. Cancer Cell, 2024, 42(7): 1268-1285.

[43] Ye F, Yuan Z, Tang Y, et al. Endocytic activation and exosomal secretion of matriptase stimulate the second wave of EGF signaling to promote skin and breast cancer invasion[J]. Cell Reports, 2024, 43(4): 114002.

[44] Batlle E, Massagué J. Transforming growth factor-β signaling in immunity and cancer[J]. Immunity, 2019, 50(4): 924-940.

[45] Liu C, Wu K, Li C, et al. SPP1+ macrophages promote head and neck squamous cell carcinoma progression by secreting TNF-α and IL-1β[J]. Journal of Experimental & Clinical Cancer Research, 2024, 43(1): 332.

（责任编辑：张　悦）