



# Identification of differentially expressed genes in lung adenocarcinoma cells using single-cell RNA sequencing not detected using traditional RNA sequencing and microarray

Zhencong Chen<sup>1</sup> · Mengnan Zhao<sup>1</sup> · Ming Li<sup>1</sup> · Qihai Sui<sup>1</sup> · Yunyi Bian<sup>1</sup> · Jiaqi Liang<sup>1</sup> · Zhengyang Hu<sup>1</sup> · Yuansheng Zheng<sup>1</sup> · Tao Lu<sup>1</sup> · Yiwei Huang<sup>1</sup> · Cheng Zhan<sup>1</sup>  · Wei Jiang<sup>1</sup> · Qun Wang<sup>1</sup> · Lijie Tan<sup>1</sup>

Received: 16 January 2020 / Revised: 28 March 2020 / Accepted: 28 March 2020 / Published online: 23 April 2020

© The Author(s), under exclusive licence to United States and Canadian Academy of Pathology 2020

## Abstract

Lung adenocarcinoma (LUAD) is the leading cause of cancer-related deaths worldwide. Traditional RNA sequencing data fails to detect the exact cellular and molecular changes in tumor cells as they make up only a small proportion of tumor tissue. 10× genomics single-cell RNA sequencing (10× scRNA-seq) and gene expression data of LUAD patients was obtained from the Department of Thoracic Surgery, Zhongshan Hospital, Fudan University, ArrayExpress, TCGA, and GEO databases. Differentially expressed genes (DEGs) were identified in LUAD and alveolar cells (DEGs-scRNA-cancer\_cell), tumor- and normal tissue-derived cells (DEGs-scRNA-sample), and normal and LUAD patients (DEGs-Bulk). Flow cytometry and qRT-PCR were performed to validate the significantly differentially expressed ligand–receptor pairs. We selected 159,219 cells and 594 samples in the scRNA-seq data and traditional RNA sequencing, respectively. A total of 1042 DEGs-scRNA-cancer\_cell, 788 DEGs-scRNA-sample, and 2510 DEGs-Bulk were identified in this study. We also identified 57 DEGs that were only detected in DEGs-scRNA-cancer\_cell (only-DEGs-scRNA-cancer\_cell). To explore the relationship between only-DEGs-scRNA-cancer\_cell and survival in LUAD, 14 and 22 only-DEGs-scRNA-cancer\_cell, which were closely related with survival in TCGA and GEO cohorts were identified. Functional enrichment analyses showed these DEGs-scRNA-cancer\_cells were mainly related to cell proliferation and immunoregulation. Our study detected and compared DEGs at different levels and revealed genes that may regulate tumor development. Our results provide a potential new protocol to determine the contribution of DEGs to cancer progression and to help identify potential therapeutic targets.

## Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide and is responsible for more than 1,700,000 new cases every year [1, 2]. Lung adenocarcinoma (LUAD),

which accounts for the majority of all lung cancers, is one of the most important subtypes of lung cancer. In recent decades, most studies on the expression profiles of LUAD patients have been based on traditional RNA-sequencing (RNA-seq) data and have improved our understanding of the occurrence and development of tumors. However, traditional RNA-seq technology is mainly focused on the “average” expression of all cells in a sample; in addition to tumor cells, tumor tissues also contain a large number of other cell types, such as macrophages and epithelial cells [3, 4]. Therefore, traditional RNA-seq technology may fail to detect the exact cellular and molecular changes in tumor cells. Many genes that may be differentially expressed and play a significant role in LUAD cannot be identified by traditional RNA-seq as abnormalities could be masked by the gene expression of other cell types. Recently, single-cell RNA-seq as an innovative technology has been used to investigate the transcriptome of different cell types [5].

These authors contributed equally: Zhencong Chen, Mengnan Zhao, Ming Li

**Supplementary information** The online version of this article (<https://doi.org/10.1038/s41374-020-0428-1>) contains supplementary material, which is available to authorized users.

✉ Cheng Zhan  
czhan10@fudan.edu.cn

✉ Wei Jiang  
jiang.wei1@zs-hospital.sh.cn

<sup>1</sup> Department of Thoracic Surgery, Zhongshan Hospital, Fudan University, No. 180, Fenglin Road, 200032 Shanghai, China

In this study, both traditional RNA-seq and 10× scRNA-seq data of LUAD patients were used to explore genes that exhibited significant differences and functions in tumor cells but were not detected in traditional RNA technology. We believe our results will improve our understanding of cellular and molecular differences between LUAD and nonmalignant tissue and provide tumor markers as well as potential therapeutic targets.

## Materials and methods

### Datasets

Nine LUAD patients from our hospital were included in scRNA-seq analyses. Twenty normal and LUAD samples were selected for flow cytometry and quantitative real-time polymerase chain reaction (qRT-PCR) analyses. We also downloaded 10× scRNA-seq data from the ArrayExpress database (<https://www.ebi.ac.uk/arrayexpress>) with accession numbers E-MTAB-6149 and E-MTAB-6653, respectively (squamous and other lung cancer subtypes were excluded in our study). Traditional RNA-seq data of LUAD was collected from TCGA (<https://tcgadata.nci.nih.gov/>) and Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) (GSE30219, GSE31210, GSE3141, GSE37745, GSE50081, and GSE68465).

### Preparation of single-cell suspensions

Tumor and nonmalignant samples were divided into single-cell suspensions by combining mechanical dissociation with enzymatic degradation of the extracellular matrix. According to the manufacturer's instructions, samples were enzymatically digested using a Tumor Dissociation Kit (Miltenyi Biotec, Gladbach, Germany). Briefly, tissues were cut into small pieces, which were mixed with 200 µL enzyme H, 25 µL enzyme A, 100 µL enzyme R, and 4.7 mL Dulbecco's Modified Essential Medium in MACS C Tube (130-094-392, Miltenyi Biotec). In the mechanical dissociation steps, we used the gentle MACS™ Dissociator (130-093-235, Miltenyi Biotec) three times and the samples were incubated for 30 min at 37 °C at each dissociation step interval. To remove large particles, erythrocytes, and dead cells, a filter (40 µm), Red Blood Cell Lysis Solution (10×) (Sigma-Aldrich, St. Louis, MO, USA), and Dead Cell Removal Kit (Miltenyi Biotec) were applied, respectively.

### Single-cell RNA seq

We prepared the single-cell RNA-seq libraries using the Chromium Single Cell 3' Reagent kit (version 2) (10×

Genomics, Pleasanton, CA, USA). Single-cell suspensions were loaded on the Chromium Single Cell Controller Instrument (10× Genomics) to generate single-cell gel beads in emulsions. Then, reverse transcription reactions used barcoded full-length cDNA followed by the disruption of emulsions using the recovery agent. To clean up cDNA, DynaBeads Myone Silane Beads (Thermo Fisher Scientific, Waltham, MA, USA) were used. Subsequently, the amplified cDNA was fragmented, end-repaired, A-tailed, index adapter ligated, and library amplified. Libraries were sequenced on the Illumina-sequencing platform (HiSeq X Ten; Illumina, San Diego, CA, USA) and 150 bp paired-end reads were generated.

### Single-cell RNA-seq data preprocessing

We used The Cell Ranger software pipeline (version 3.0.0) to demultiplex cellular barcodes and map reads to the genome. STAR aligner was applied to generate normalized aggregate data across samples and obtain a matrix of gene counts versus cells.

### CNV estimation

To estimate the initial copy number variations (CNVs) in each cell cluster, we used R package “inferCNV” to count the CNVs [6]. The input file is a matrix of the expression level of each cells without normalization, and we used cells except “Alveolar cluster” and “Cancer cluster” as background to calculate the CNVs score.”

### Statistical analysis

#### 10× scRNA-seq data analysis

The 10× scRNA-seq data analyses performed in R version 3.6.1 were as follows: (1) Seurat R package [7] was used to convert 10× scRNA-seq data as a Seurat object and the “FindVariableFeatures” function was used to select the top 2000 highly variable genes after quality control; (2) principal component analysis (PCA) was performed based on the 2000 genes to analyze the 10× scRNA-seq data; (3) Harmony R package [8] was used to integrate single-cell data and function “RunHarmony” was applied in a Seurat object to integrate the scRNA-seq from different dataset; (4) after integrating single-cell data, uniform manifold approximation and projection (UMAP) were applied to explore the 10× scRNA-seq data; and (5) SingleR package [9], CellMarker dataset [10], and previous studies were applied to recognize the different cell types.

## Identification of differentially expressed genes (DEGs)

After identifying the cell types in 10× scRNA-seq, R package MAST [11] was used to identify DEGs between LUAD and alveolar cells (DEGs-scRNA-cancer\_cell). In addition, DEGs between the tumor- and normal tissue-derived cells (DEGs-scRNA-sample) were determined using the edgeR package in R. To detect DEGs between normal and LUAD patients (DEGs-Bulk), we also performed differential expression analysis of TCGA data using the edgeR package. Moreover, to compare DEGs-scRNA-cancer\_cell, DEGs-scRNA-sample, and DEGs-Bulk, equal numbers of genes were included in the DEG analyses. The statistical threshold for significance was a false discovery rate (FDR) <0.05 and fold change >1.5.

## Functional enrichment analyses

To explore differential activities of pathways between cells derived from LUAD tumor or nonmalignant tissue, gene set variation analysis (GSVA) was performed using the GSEABase package [12]. A described curated database was also used to evaluate metabolic pathway activities [13]. The GSVA package [14] was applied to assign pathway activity scores to each cell type.

To investigate the potential molecular mechanisms of the immune-related genes, a comprehensive immune-related gene set was obtained from the Immunology Database and Analysis Portal (ImmPort) database (<https://immport.niaid.nih.gov>) [15]. This was used to identify immune genes that were differentially expressed between LUAD tumor cell and alveolar cell clusters. In addition, gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were performed by Metascape (<http://metascape.Org>) [16]. A significant difference in GO or KEGG pathways was defined as  $P < 0.01$  and number of enriched genes >3.

## Survival statistical analysis

Survival statistical analyses were performed in IBM SPSS statistics software, version 22.0 (IBM, Inc., Armonk, NY, USA) and R version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria). Kaplan–Meier and log-rank tests were used to construct and compare survival curves. To confirm whether expression of the selected gene was associated with poor survival, patients with complete follow-up records in TCGA and the GEO database were selected. We split the patients into a high-expression group (>median expression level across all samples) and a low expression group (≤median expression level across all samples). A significant difference of survival analysis was defined as  $P < 0.05$ .

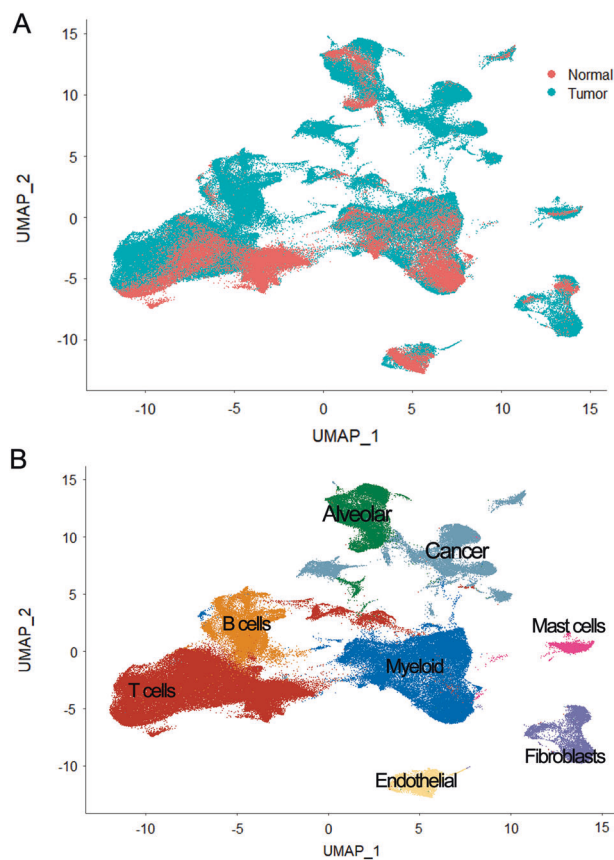
## Validation

To perform flow cytometry, we used phosphate-buffered saline with 3% fetal bovine serum and 20 µg/mL human IgG (Sigma-Aldrich) to block nonspecific antibody binding for 15 min. Then, cells were incubated with allophycocyanin-conjugated mouse anti-human EPCAM (5 µL/10<sup>6</sup> cells; cat. no.: 566658, BD Biosciences, San Jose, CA, USA), BV421-conjugated mouse anti-human CD45 (5 µL/10<sup>6</sup> cells; cat. no.: 304022, BioLegend, San Diego, CA, USA), or PE-conjugated mouse anti-human FOLR1 (10 µL/10<sup>6</sup> cells; cat. no.: FAB5646P, R&D Systems, Minneapolis, MN, USA) for 30 min on ice. To quantitate and isolate stained cells, Fortessa analyzer (BD Biosciences) and FACSaria II (BD Biosciences) were applied in our study, and FlowJo software (TreeStar, Woodburn, OR, USA) was used to generate the flow described above. In qRT-PCR analyses, sorted cells were subjected to RNA extraction and reverse transcription using a kit (Takara, Kusatsu, Japan) prior to the experiment.

## Results

### scRNA-seq and cell typing of nonmalignant and LUAD lung samples

In 10× scRNA-seq data analysis, after quality filtering, as shown in Supplementary Fig. 1, before we integrated the scRNA-seq from different dataset, cells mainly clustered by dataset, after integration, cells were mixed together, which suggested 159,219 cells derived from three databases were well integrated. Of these, 50,845 (31.9%) cells were from the E-MTAB-6149 database, 36,143 (22.7%) cells were from the E-MTAB-6653 database, and 72,231 (45.4%) cells were from our department (Supplementary Fig. 1). As shown in Fig. 1a, 122,082 cells (76.7%) originated from 18 malignant lung tissue samples and 37,137 (23.3%) originated from seven normal lung tissue samples (Supplementary Tables 1 and 2). After applying PCA and uniform manifold approximation and projection analysis on highly variable genes ( $n = 2000$ ) expressed across all 159,219 cells, eight distinct groups of cells were identified Fig. 1b. We assigned these cell groups to known cell types based on expression of the following known markers: CD3D, TRAC, TRBC2 (T cells), C1QB, LYZ, CD68 (myeloid cells), CLDN5, FCN3, RAMP2 (endothelial cells), C1R, COL1A1, DCN (fibroblasts), CPA3, TPSAB1 (mast cells), TPSB2, CD79A, IGHG3, and IGKC (B cells) [9, 17, 18]. In addition, to better identify the malignant cells and normal lung cells, we first we mapped the expression of six marker genes (EPCAM, MDK, and SOX4 for cancer cells, and EPCAM, AGR3, FOLR1, and SFTPD for normal epithelial



**Fig. 1 Overview of the 1,159,219 single cells from eighteen tumor samples and seven normal samples. a** The sample origin of the cells; **b** The cell types identified by marker genes.

cells) to distinguish the two types of cells in our study. As Fig. 2 shows, the expression of EPCAM, MDK, and SOX4 is elevated in cancer cell cluster, while the expression of EPCAM, AGR3, FOLR1, and SFTPD is elevated in normal epithelial cells cluster. Next, according to the origin of cells, we noticed that cells which both had high expression levels of normal epithelial marker genes and derived from normal tissues were mainly enriched in “Alveolar cluster”, which suggested that the “Alveolar cluster” may be normal lung cells. In addition, previous studies had demonstrated that compared with normal cells, tumor cells had higher CNV levels. Therefore, to further validate the malignant cells, we estimated the CNV in “Cancer cluster” and “Alveolar cluster” respectively. Our results revealed that compared with “Alveolar cluster”, the cluster which we identified as malignant cells exhibited remarkably higher CNV levels, suggesting the reasonable cell annotation in our study (Supplementary Fig. 2). Detailed information of the marker genes for each cell type is shown in Fig. 2 and Supplementary Fig. 3.

To better characterize the heterogeneity of cancer cells, we regrouped these cancer cells and results revealed the cancer cells were reclustered to three subclusters

(Supplementary Fig. 4). Next we attempted to identify the marker genes for each subcluster; furthermore, to investigate the biological functions of each cluster, enrichment analyses were also performed in our study. As shown in Supplementary Fig. 5, Subcancer1 were more related to the cell metabolism and cell metabolism, while Subcancer2 and Subcancer3 were more closely to cell migration and the immune response respectively. Taken together, our results revealed the huge heterogeneity in tumor cells.

Next we focused on the DEGs-scRNA analyses. A total of 13,749 genes, present in both 10× scRNA-seq data and traditional RNA-seq data, were selected for DEG analyses. In DEGs-scRNA-cancer\_cell analyses, 32,960 cells (18,248 LUAD and 14,712 alveolar cells) were included in our study. Ultimately, 1042 DEGs-scRNA-cancer\_cell were identified between LUAD and alveolar cells, and 808 upregulated and 234 downregulated genes in the cancer group were identified (Supplementary Table 3). Moreover, compared with nonmalignant and malignant tissue-derived cells, 788 DEGs-scRNA-sample (72 upregulated and 716 downregulated genes) were detected in this study (Supplementary Table 3).

### Identification of DEGs in traditional RNA-seq

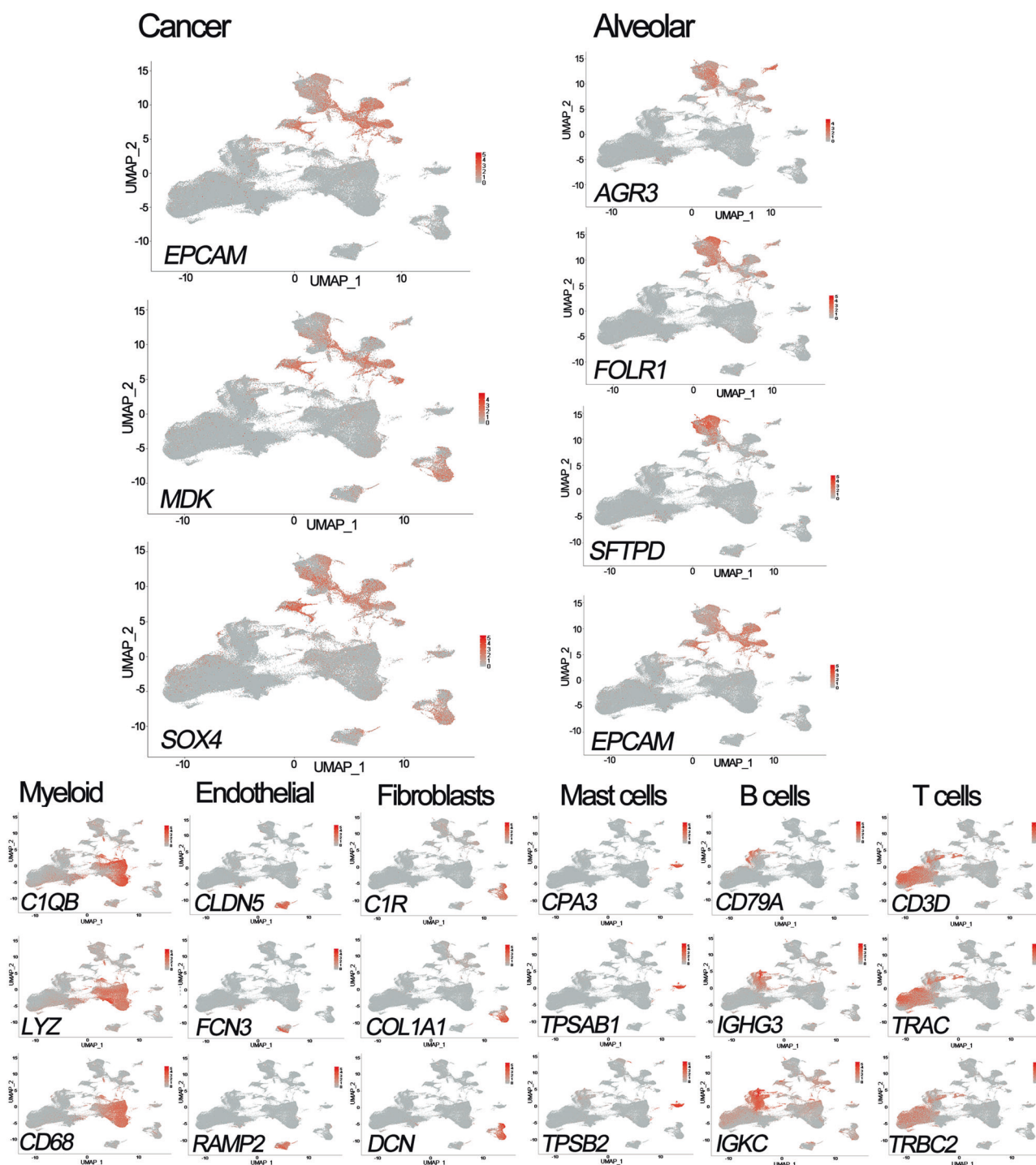
Based on the TCGA database, we obtained 535 tumor and 59 normal samples in our DEGs-Bulk analysis. The DEGs-Bulk were selected with strict criteria of fold change >1.5 and FDR <0.05. A total of 2510 genes were identified: 1266 and 1244 were significantly upregulated and downregulated, respectively, in the LUAD samples (Supplementary Table 3).

### Compared DEGs among DEGs-scRNA-cancer\_cell, DEGs-scRNA-sample and DEGs-Bulk

Intriguingly, there were 245 and 197 DEGs-scRNA-cancer\_cell that were also detected in DEGs-scRNA-sample and DEGs-Bulk analyses, respectively; a total of 57 genes were only found in DEGs-scRNA-cancer\_cell (only-DEGs-scRNA-cancer\_cell), including 51 upregulated and 6 downregulated genes. The top eight highly variable only-DEGs-scRNA-cancer\_cell were PDCD6, HMGA1, CYBA, SPAG7, UQCERS1, FAM20A, PDCD4, and EMC6. The top 30 highly variable genes were shown in Fig. 3 and Supplementary Fig. 6.

A total of 12,164, 10,349, and 9964 non-DEGs-scRNA-cancer\_cell were also confirmed as non-DEGs-scRNA-sample, non-DEGs-Bulk, and non-DEGs-scRNA-sample&non-DEGs-Bulk, respectively. A total of 543 and 2313 non-DEGs-scRNA-cancer\_cell were also identified as DEGs-scRNA-sample and DEGs-Bulk. Notably, we found that these DEGs-scRNA-sample&non-DEGs-scRNA-cancer\_cell

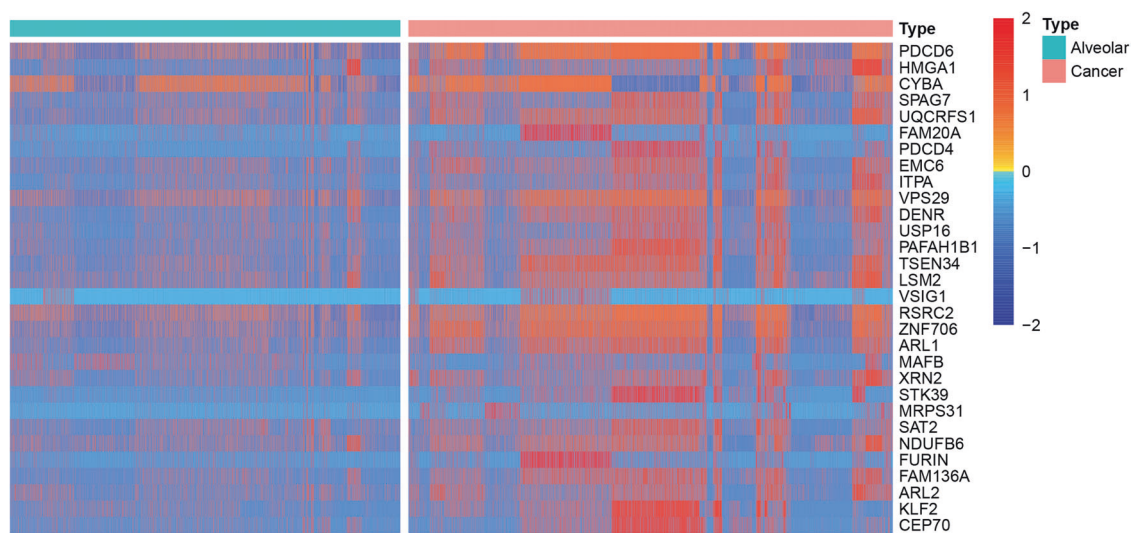




**Fig. 2** UMAP plot of the marker genes. Expression of marker genes for Cancer, Alveolar, Myeloid cells, Endothelial cells, Fibroblasts, Mast cells, B cells, and T cells.

or DEGs-Bulk&non-DEGs-scRNA-cancer\_cell had relatively high expression levels in non-LUAD tumor cell clusters compared with LUAD tumor cell clusters, such as C1QA (DEGs-scRNA-sample and non-DEGs-scRNA-cancer\_cell) and RETN (DEGs-Bulk and non-DEGs-scRNA-cancer\_cell) in Myeloid

cluster (Supplementary Fig. 7). This suggested that gene expression of non-LUAD tumor cells has a significant effect on differential analysis that could result in the inability to detect these genes in the DEGs-scRNA-sample or DEGs-Bulk analyses.



**Fig. 3 Heatmap of DEGs for LUAD and alveolar cells.** Heatmap of the top 30 only-DEGs-scRNA-cancer\_cell between LUAD and alveolar cells.

### Survival analyses

To explore the relationship between only-DEGs-scRNA-cancer\_cell and survival in LUAD, we performed survival analyses in only-DEGs-scRNA-cancer\_cell. As a result, 492 and 1063 LUAD patients with complete follow-up records were included from TCGA and GEO databases. We respectively obtained 14 and 22 only-DEGs-scRNA-cancer\_cell, which were closely related with survival in the TCGA and GEO cohorts. Moreover, 8 only-DEGs-scRNA-cancer\_cell, which had a statistically significant influence on the prognosis, were identified in TCGA and GEO survival analyses (Fig. 4).

### Functional enrichment analyses

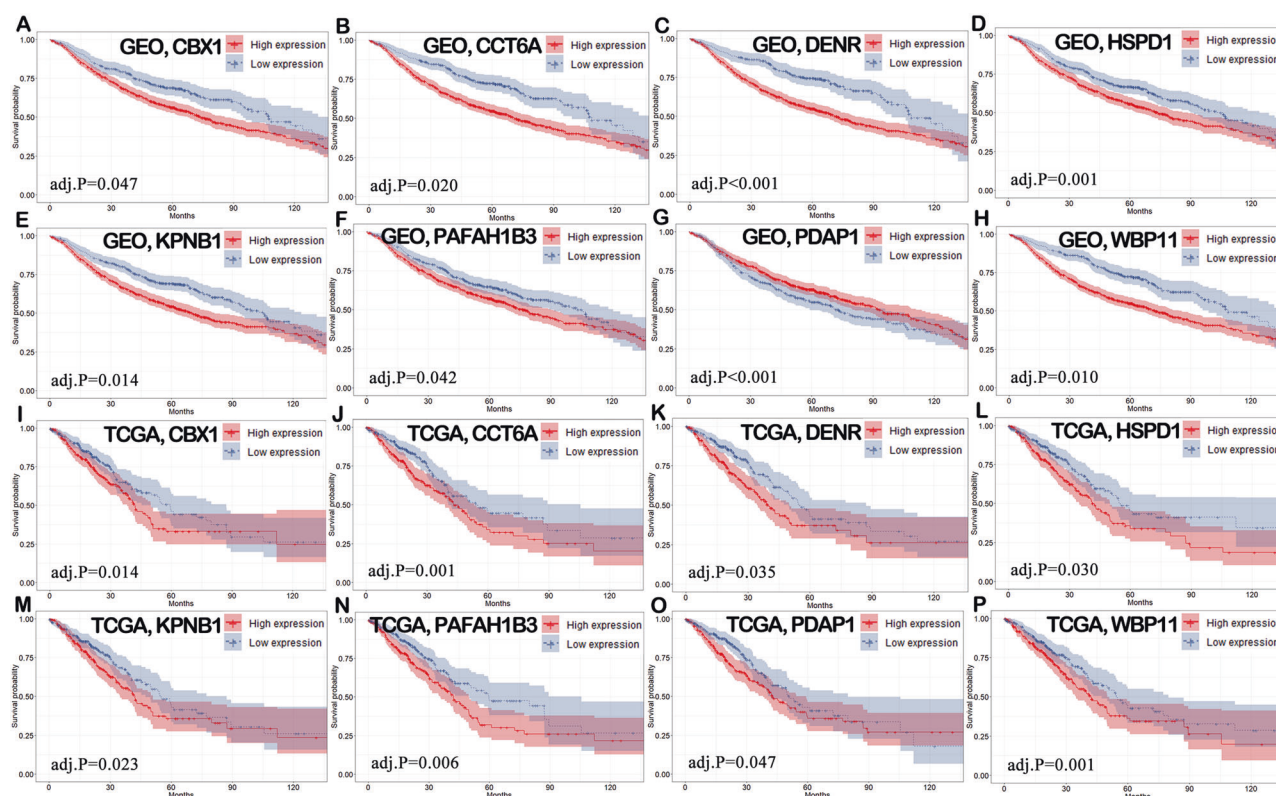
A direct comparison of LUAD tumor and alveolar cells was performed based on DEGs-scRNA-cancer\_cell. As shown in Fig. 5, Myc targets v2, G2M checkpoint, E2F targets, glycolysis, and DNA repair were the top five enriched pathways in LUAD tumor cells. The glycolysis pathway as a candidate target for combinatorial therapeutic intervention has been identified in several studies. Recent studies showed that tumor glycolysis was associated with immune resistance in lung cancer, and glycolysis-related genes were upregulated in lung cancer patient samples poorly infiltrated by T cells [19]. Overexpression of glycolysis-related molecules impaired T cell killing of tumor cells, whereas inhibition of glycolysis enhanced T cell-mediated antitumor immunity in vitro and in vivo. E2F transcription factors (E2Fs) have a clearly defined role in cell-cycle control, while recent work has revealed that E2Fs play key roles in mediating tumor development and metastasis [20]. Recent

studies have suggested that the E2F pathway controls the expression of genes important for angiogenesis, remodeling of the extracellular matrix, tumor cell survival, and tumor cell interactions with vascular endothelial cells that facilitate metastasis to the lungs [21]. We also found that DEGs in LUAD tumor cells were enriched in G2M checkpoint and DNA repair pathways and played an important role in cell-cycle regulation.

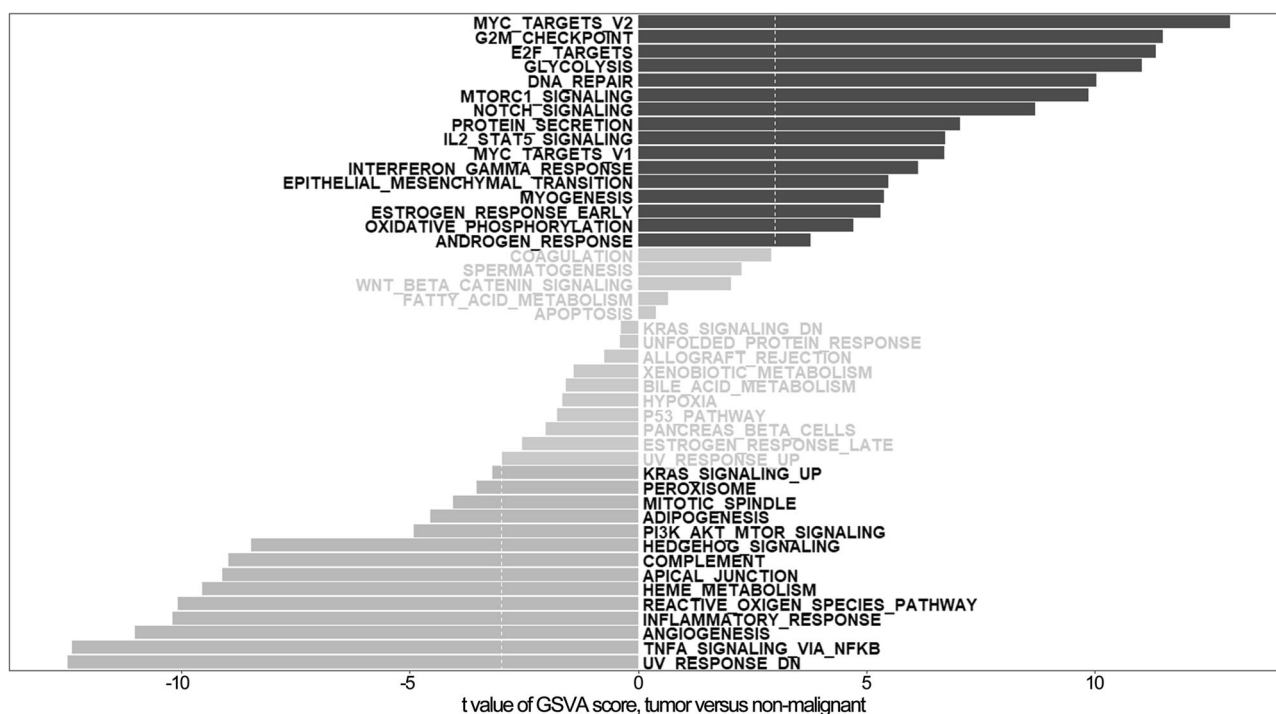
To explore the relationship between the DEGs-scRNA-cancer\_cell expression and immunotherapy in LUAD, the immune-related genes that were only identified in DEGs-scRNA-cancer\_cell were selected by the ImmPort database. Functional enrichment analyses were performed as shown in Fig. 6. Immune-related genes were mainly enriched in regulation of cytokine production, response to peptide (GO), and cytokine–cytokine receptor interaction.

### Flow cytometry and qRT-PCR analyses

Flow cytometry was performed to validate tumor cells marked with EPCAM, and alveolar cells marked with FOLR1 in normal and LUAD samples (Fig. 7). To validate the DEGs-scRNA-cancer\_cell in our study, qRT-PCR was also performed. As shown in Fig. 8, we observed that the expression levels of PDCD6 ( $P < 0.01$ ), HMGA1 ( $P < 0.01$ ), SPAG7 ( $P < 0.01$ ), UQCRCF1 ( $P < 0.01$ ), PDCD4 ( $P < 0.01$ ), and EMC6 ( $P < 0.01$ ) were significantly increased in LUAD tumor cells and the expression levels of FAM20A ( $P < 0.01$ ) and CYBA ( $P < 0.01$ ) were increased in alveolar cells. These findings were consistent with the DEG-group analyses, suggesting that 10× scRNA-seq data analysis can effectively explore DEGs between LUAD tumor cells and alveolar cells.

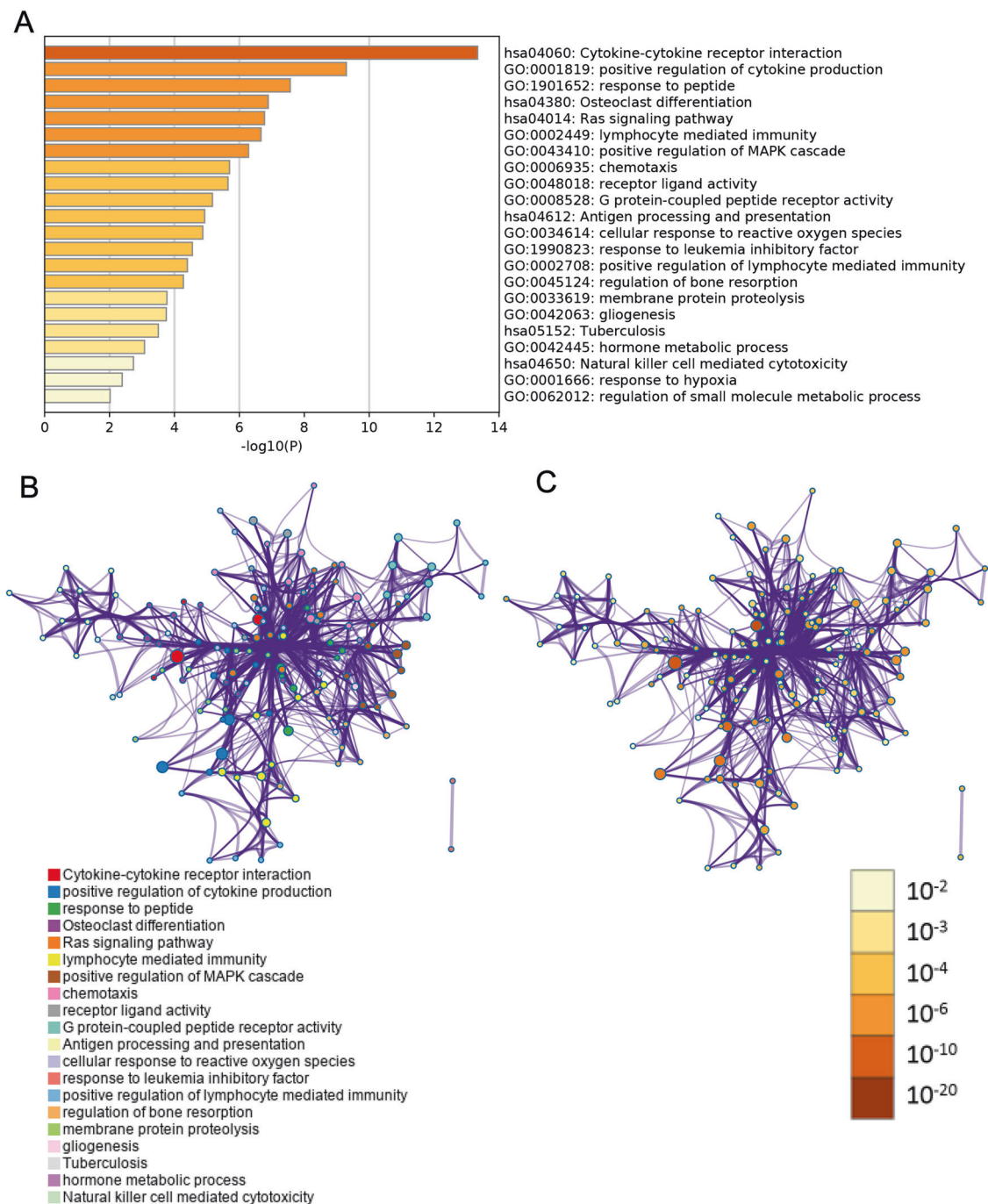


**Fig. 4** Kaplan–Meyer plot of OS in TCGA and GEO databases with high or low expression group of 8 only-DEGs-scRNA-cancer\_cell. The *P* values of a–p were all less than 0.05.



**Fig. 5** GSVA analysis in LUAD tumor and alveolar cells. GSVA analysis of the hallmark pathways between LUAD tumor and alveolar cells.





**Fig. 6 Functional analysis of only-DEGs-scRNA-cancer\_cell related immune genes.** **a** Heatmap of enriched terms across input gene lists, colored by *P* values. Network of enriched terms: **b** colored by

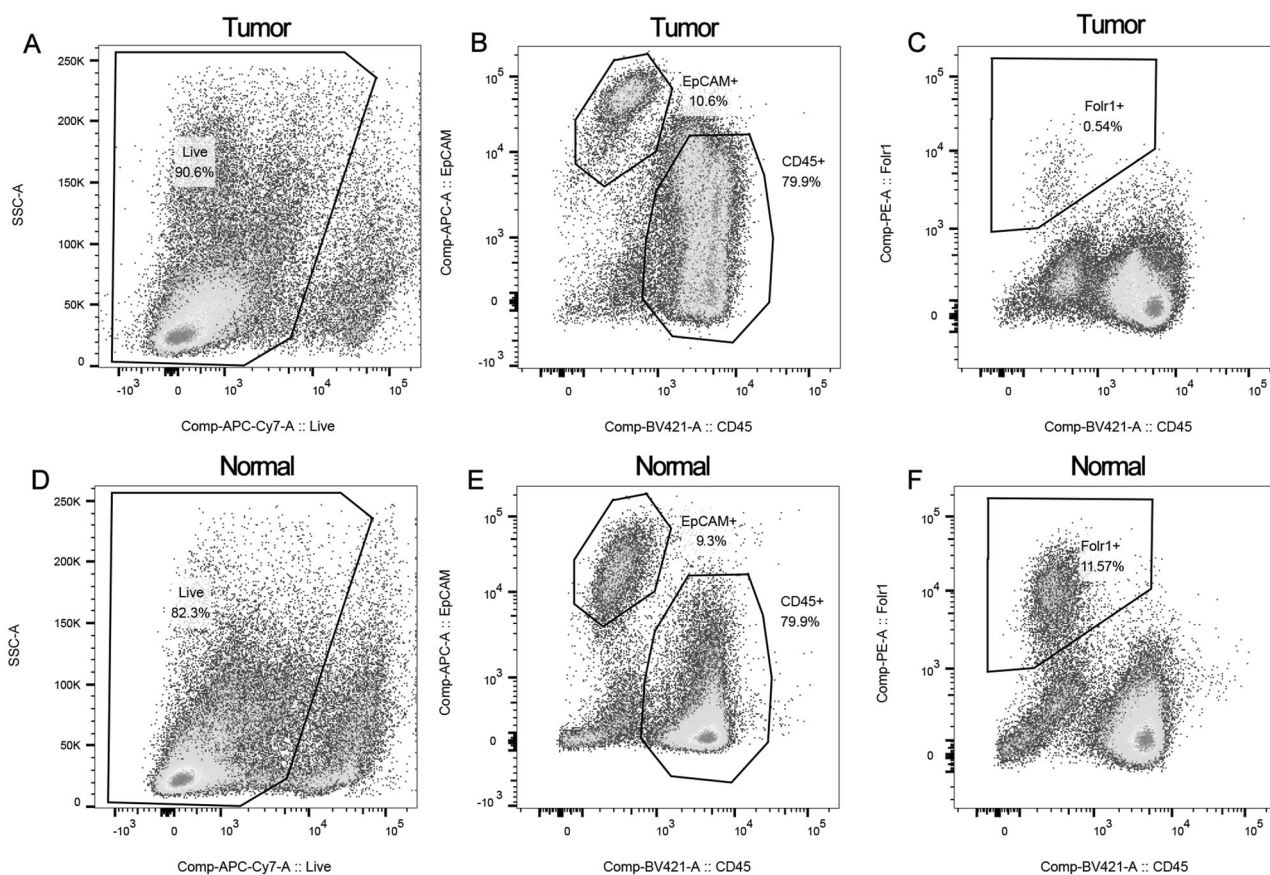
cluster ID, where nodes that share the same cluster ID are typically close to each other; **c** colored by *P* value, where terms containing more genes tend to have a more significant *P* value.

## Discussion

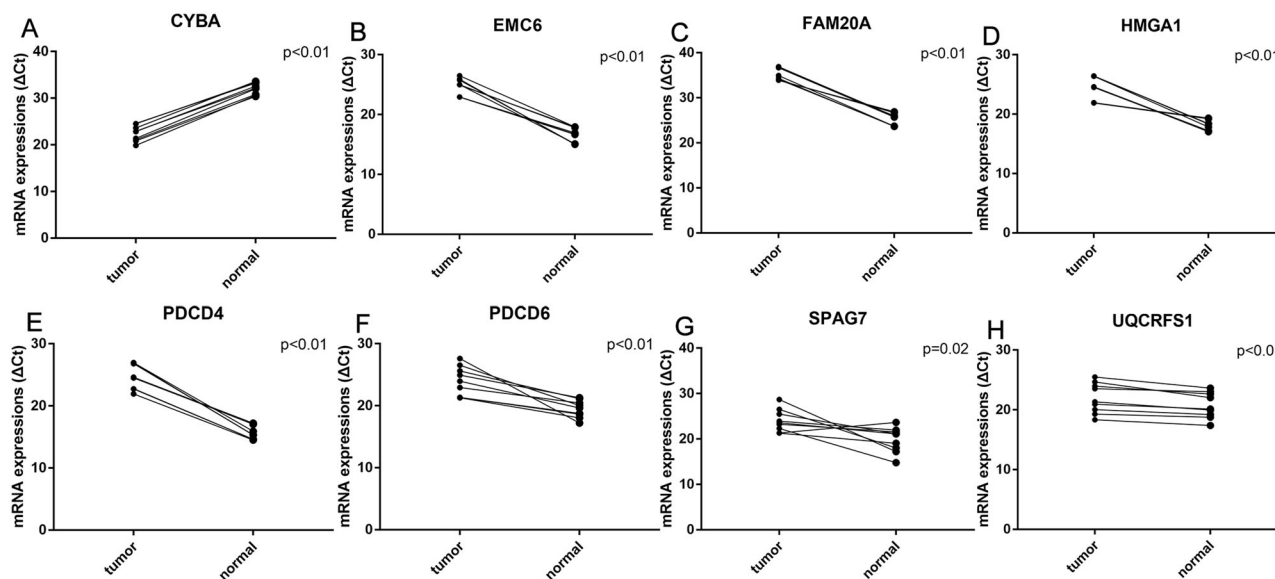
This study was based on integrated analysis of 10× scRNA-seq data, which included more than 20 samples from the Department of Thoracic Surgery, Zhongshan Hospital, Fudan University, ArrayExpress. A total of 159,219 cells were included in our study. Of these, 122,082 cells (76.7%)

originated from LUAD tissues and 37,137 (23.3%) originated from non-malignant lung tissues. In our analysis, we found that the expression of EPCAM was elevated in both cancer cells and epithelial cells. However, we used the expression levels of tumor marker genes, the origin of cells, and estimated value of CNVs to distinguish the two types of cells as suggested in previous studies. By analyzing





**Fig. 7** Flow cytometry for cancer cells and alveolar cells. Identified and sorted the cancer cells and alveolar cells in tumor sample and normal sample by flow cytometry.



**Fig. 8** Validation of the top eight highly variable only-DEGs-scRNA-cancer\_cell. The mRNA relative expression level of PDCD6 ( $P < 0.01$ ), HMGA1 ( $P < 0.01$ ), SPAG7 ( $P = 0.02$ ), UQCERS1

( $P < 0.01$ ), PDCD4 ( $P < 0.01$ ), EMC6 ( $P < 0.01$ ), and FAM20A ( $P < 0.01$ ) were significantly increased in LUAD tumor cells while the expression levels of CYBA ( $P < 0.01$ ) were increased in alveolar cells.

hundreds of thousands of cells from LUAD and normal tissues, we identified hundreds of genes whose expression was significantly altered in LUAD tumor cells that could not be distinguished by RNA-seq or microarray. A large proportion of these genes was correlated with the prognosis of LUAD, which improves our understanding of the occurrence and development of LUAD and will help in the development of therapeutic targets.

In the present study, we identified 51 genes significantly altered in cancer cells compared with alveolar cells, which could not be identified using traditional RNA-seq or microarray. Interestingly, we found that based on the top eight highly variable only-DEGs-scRNA-cancer\_cell, the high expression levels of HMGA1 and EMC6 were related to poorer prognosis in both the TCGA and GEO databases. It has been suggested that HMGA1 can regulate gene expression and alter chromatin structure and serves as an architectural transcription factor [22]. Recent studies have revealed that HMGA1 is a promising prognostic biomarker for several cancer types, including lung cancer because of its association with poor differentiation status [23]. Further analyses showed that HMGA1 was significantly associated with CD8+ and CD4+ T cells in many carcinomas [24]. EMC6, also called transmembrane protein 93 (TMEM93), plays an important role in autophagy and is located on chromosome 17p13.2 [25, 26]. Previous studies have shown that EMC6 is an autophagy-related gene that contains two conserved transmembrane domains, and interacts with Beclin1 and Rab5a [27]. Shen et al. reported [28] that EMC6-mediated autophagy was associated with inactivation of the PIK3CA/AKT/mTOR signaling pathway. Although previous studies have shown that EMC6 is widely expressed in normal human tissues [27], Shen et al. [25] suggested that the expression level of EMC6 in cancer samples is significantly higher than that in normal tissues. The expression and role of EMC6 in tumor progression has not yet been fully investigated; further studies are necessary.

Tumors are mixtures of different compartments which include immune, stroma, tumor, and normal cells [29]. The tumor microenvironment (TME) is an important component of tumor tissues, which plays an active role in cancer progression and therapeutic responses [3, 30]. For instance, several studies have suggested that T cells, which are closely related to immune therapy and patient survival, represent the most prevalent cell type in the TME of LUAD [17, 31]. To study the TME using traditional RNA gene expression, many computational algorithms have been developed to deconvolve the mixed signals in bulk cancer tissue, such as ESTIMATE [32] and CIBERSORT [33]. However, there are some limitations in present traditional RNA gene expression-based deconvolution methods. For example, CIBERSORT requires information about the pure expression of compartments as a reference, which is

difficult to obtain. Moreover, true compartments in each sample are usually uncertain and samples are inherently heterogeneous, which may affect the accuracy of calculations [34]. Using 10× scRNA-seq data, we can identify and characterize stromal cells and the extracellular matrix, which will help us to improve our understanding of the contribution of the TME to cancer progression and potential targets for immunotherapy in the future.

Heterogeneity, which contributes to the existence of different compartments within a tumor, is well known as a prominent characteristic of tumors [35]. In addition, it is essential to investigate the evolution of these compartments since it represents a driving force behind tumor development, tumor progression, drug resistance, and metastasis [36]. Traditional RNA-seq provides an average of gene expression in the sample, and the analysis of tumor gene expression can be obscured by the existence of non-neoplastic cells. Compared with RNA-seq technologies, 10× scRNA-seq allows us to delve into the transcriptome of individual cells and enables exploration tumor heterogeneity [5, 37]. In addition, we found that T cells are the dominant cell type in tumor and normal samples, indicating that the gene expression of T cells has a significant effect on overall gene expression, consistent with previous studies [38, 39].

In summary, the 10× genomics technique of single-cell RNA seq has some limitations, such as only sequencing the 3' end, expensive cost, and relatively low coverage [34]. All single-cell RNA-seq technologies, including 10× scRNA-seq technology, are flawed in reliable sensitivity of gene detection. Furthermore, the expression levels of genes detected in cells may not reflect the true expression levels, since the zero inflated exists in most scRNA-seq technologies. However, when compared with traditional RNA-seq technology, scRNA-seq had the relatively good performance in distinguishing the molecular characteristics in each cell type [40]. Using 10× scRNA-seq data, we detected and compared DEGs from different levels. We revealed the critical genes that may regulate tumor proliferation, invasion, metastasis, and drug resistance in LUAD. In addition, our data can be a valuable resource for people to further explore the biological insights of LUAD in future, especially in the analysis of tumor immune phenotypic. We believe these observations will improve our understanding of the contribution of the DEGs to cancer progression and potential therapeutic targets.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (No. 81401875), Research Development Fund of Zhongshan Hospital, Fudan University (2019ZSFZ002 and 2019ZSFZ19), and Zhengyi Scholar Foundation of School of Basic Medical Sciences, Fudan University (Grant No. S22-11). We would like to thank International Science Editing Co. for the language editing service.

**Author contributions** CZ, WJ, and ZC conceived the study. ZC, MZ, and ML performed most of the bioinformatics analysis and wrote the manuscript. QS, YB, and JL collected the tumor samples and analyzed the data. ZH, YZ, and TL performed the validation of this study. YH and QW helped project design and manuscript editing. LT supervised this study.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394–424.
- Siegel RL, Miller KD, Jemal A. Cancer statistics. *CA Cancer J Clin*. 2019;69:7–34.
- Guo JY, Hsu HS, Tyan SW, Li FY, Shew JY, Lee WH, et al. Serglycin in tumor microenvironment promotes non-small cell lung cancer aggressiveness in a CD44-dependent manner. *Oncogene*. 2017;36:2457–71.
- Su S, Chen J, Yao H, Liu J, Yu S, Lao L, et al. CD10+GPR77+ cancer-associated fibroblasts promote cancer formation and chemoresistance by sustaining cancer stemness. *Cell*. 2018;172:841–56.
- Olsen TK, Baryawno N. Introduction to single-cell rna sequencing. *Curr Protoc Mol Biol*. 2018;122:e57.
- Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*. 2014;344:1396–401.
- Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 2015;161:1202–14.
- Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16:1289–96.
- Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*. 2019;20:163–72.
- Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res*. 2019;47:D721–d8.
- Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. 2015;16:278.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102:15545–50.
- Gaude E, Frezza C. Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. *Nat Commun*. 2016;7:13041.
- Hanzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform*. 2013;14:7.
- Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, et al. ImmPort: disseminating data to the public for the future of immunology. *Immunol Res*. 2014;58:234–9.
- Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10:1523.
- Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med*. 2018;24:1277–89.
- Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res*. 2018;47:D721–D8.
- Cascone T, McKenzie JA, Mbofung RM, Punt S, Wang Z, Xu C, et al. Increased tumor glycolysis characterizes immune resistance to adoptive T cell therapy. *Cell Metab*. 2018;27:977–87 e4.
- Nevins JR. The Rb/E2F pathway and cancer. *Hum Mol Genet*. 2001;10:699–703.
- Johnson J, Thijssen B, McDermott U, Garnett M, Wessels LF, Bernards R. Targeting the RB-E2F pathway in breast cancer. *Oncogene*. 2016;35:4829–35.
- Frasca F, Rustighi A, Malaguarnera R, Altamura S, Vigneri P, Del Sal G, et al. HMGA1 inhibits the function of p53 family members in thyroid cancer cells. *Cancer Res*. 2006;66:2980–9.
- Hassan F, Ni S, Arnett TC, McKell MC, Kennedy MA. Adenovirus-mediated delivery of decoy hyper binding sites targeting oncogenic HMGA1 reduces pancreatic and liver cancer cell viability. *Mol Ther Oncolytics*. 2018;8:52–61.
- Vallacchi V, Vergani E, Camisaschi C, Deho P, Cabras AD, Sensi M, et al. Transcriptional profiling of melanoma sentinel nodes identify patients with poor outcome and reveal an association of CD30(+) T lymphocytes with progression. *Cancer Res*. 2014;74:130–40.
- Shen M-X, Ding J-B. Expression levels and roles of EMC-6, Beclin1, and Rab5a in the cervical cancer. *Eur Rev Med Pharmacol Sci*. 2017;21:3038–46.
- Li Y, Zhao Y, Hu J, Xiao J, Qu L, Wang Z, et al. A novel ER-localized transmembrane protein, EMC6, interacts with RAB5A and regulates cell autophagy. *Autophagy*. 2013;9:150–63.
- White E. Deconvoluting the context-dependent role for autophagy in cancer. *Nat Rev Cancer*. 2012;12:401–10.
- Shen X, Kan S, Hu J, Li M, Lu G, Zhang M, et al. EMC6/TMEM93 suppresses glioblastoma proliferation by modulating autophagy. *Cell Death Dis*. 2016;7:e2043.
- Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer*. 2012;12:323–34.
- Katsumata S, Aokage K, Miyoshi T, Tane K, Nakamura H, Sugano M, et al. Differences of tumor microenvironment between stage I lepidic-positive and lepidic-negative lung adenocarcinomas. *J Thoracic and Cardiovas Surg*. 2018;156:1679–88.e2.
- Thommen DS, Koelzer VH, Herzig P, Roller A, Trefny M, Dimeloe S, et al. A transcriptionally and functionally distinct PD-1(+) CD8(+) T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. *Nat Med*. 2018;24:994–1004.
- Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;4:2612.
- Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12:453–7.
- Peng XL, Moffitt RA, Torphy RJ, Volmar KE, Yeh JJ. De novo compartment deconvolution and weight estimation of tumor samples using DECODER. *Nat Commun*. 2019;10:4729.

35. Yadav VK, De S. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Brief Bioinform.* 2015;16:232–41.
36. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell.* 2018;174:1293–308. e36.
37. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. *Mol Cell.* 2015;58:610–20.
38. Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat Med.* 2018;24:978–85.
39. Savas P, Virassamy B, Ye C, Salim A, Mintoff CP, Caramia F, et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat Med.* 2018;24:986–93.
40. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* 2019;29:725–38.