Supplementary materials

Methods

Patient selection

The study was approved by the Research Ethics Committee of Guangdong Provincial People's Hospital (No. GDREC2019304H(R1)). All patients provided written informed consent for participation; separate consent was obtained for tumor specimens for biomarker analyses. Tumor tissues were freshly obtained from patients undergoing lung resection at the hospital. Nine samples were collected from eight patients (Supplementary Table 1). The eligibility criteria were: confirmed lung adenocarcinoma, treatment naïve, adequate surgical specimens of > 1 cm diameter, without other cancer. After resection, tumor-tissue samples were collected and immediately transferred for tissue preparation. Half of the tissues were subjected to single-cell isolation and the other half were subjected to nested polymerase chain reaction (nested-PCR) and mIHC for analysis of EGFR mutations and interested protein expression, respectively. The data of three samples from one patient with early stage multiple primary lung cancer, showing remarkable tumor shrinkage in a nodule and no response in two other nodules after treatment with three cycles of neoadjuvant pembrolizumab, were analyzed in this study [7]. The dataset accession number is GSE146100 (Zhang et al.).

EGFR mutation analysis by direct sequencing

All nested-PCR amplified products that were positive by agarose gel electrophoresis

Yang L, *et al. J Immunother Cancer* 2022; 10:e003534. doi: 10.1136/jitc-2021-003534

were sequenced to determine the EGFR status. The products were purified, labeled using a Big-Dye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA), and sequenced using an ABI 3100 Genetic Analyzer (Applied Biosystems). The sequencing reactions were confirmed by two experienced, independent readers.

Single-cell dissociation

Tissues were surgically removed and placed in MACS Tissue Storage Solution (Miltenyi Biotec) until processing. Tissue samples were processed as described below. Briefly, samples were first washed with phosphate-buffered saline (PBS), minced into small pieces (~ 1 mm^3) on ice, and enzymatically digested with 250 U/mL collagenase I (Gibco), 100 U/mL collagenase IV (Gibco), and 30 U/mL DNase I (Worthington) for 45 min at 37°C, with agitation. Next, samples were passed through a 70-µm cell strainer and centrifuged at 300 g for 5 min. The supernatant was decanted, and the pelleted cells were suspended in Red Blood Cell Lysis Buffer (Miltenyi Biotec). After washing with PBS containing 0.04% BSA, cell pellets were resuspended in PBS containing 0.04% BSA and passed through a 35-µm cell strainer. To assess their viability, dissociated single cells were stained using Calcein-AM (Thermo Fisher Scientific) and Draq7 (BD Biosciences).

Single-cell RNA sequencing

Single-cell RNA-seq experiment was performed by NovelBio Bio-Pharm Technology

Co., Ltd. The BD Rhapsody system was used to capture the transcriptomic information of the nine sample-derived single cells. Single-cell capture was achieved by random distribution of a single-cell suspension across > 200,000 microwells by a limiting-dilution approach. Beads with oligonucleotide barcodes were added to saturation so that a bead was paired with each cell. The cells were lysed in the microwell to hybridize mRNA molecules to barcoded capture oligos on the beads. Beads were collected in a single tube for reverse transcription and *Exo*I digestion. Each cDNA was tagged at the 5′-end (*i.e.*, the 3′-end of an mRNA transcript) with a unique molecular identifier (UMI) and cell barcode indicating its cell of origin. Whole-transcriptome libraries were prepared using the BD Rhapsody single-cell whole-transcriptome amplification (WTA) workflow, including random priming and extension (RPE), RPE amplification PCR, and WTA index PCR. The libraries were quantified using a High-Sensitivity DNA Chip (Agilent) on a Bioanalyzer 2200 and the Qubit High-Sensitivity DNA assay (Thermo Fisher Scientific). Sequencing was performed using an Illumina sequencer (San Diego, CA, USA) on a 150-bp paired-end run. Single-cell RNA sequencing datasets generated in this study are available on the GEO database under the accession number GSE171145.

Single-cell RNA analysis

scRNA-seq data analysis was performed by NovelBio Bio-Pharm Technology Co., Ltd. On the NovelBrain Cloud Analysis Platform. We used fastp[1] with default parameter filtering of the adaptor sequences and removed low-quality reads. UMI-tools was

used for single-cell transcriptome analysis to identify the cell barcode whitelist. UMI-based clean data were mapped to the human genome (Ensemble version 91) by STAR[2] mapping with customized parameters from the UMI-tools standard pipeline to calculate the UMI counts. Cells with > 200 expressed genes and a mitochondria UMI rate of < 40% passed the cell-quality filtering, and mitochondrial genes were removed from the expression table. The Seurat package (version 3.1.4, https://satijalab.org/seurat/) was used for cell normalization and regression based on the expression table according to the UMI counts and mitochondrial rates to obtain scaled data. Principal component analysis (PCA) was performed based on the scaled data with the top 2,000 highly variable genes, and the top 10 principals were used for tSNE and UMAP construction.

Using the graph-based cluster method (resolution 0.8), we acquired the unsupervised cell cluster results and determined marker genes using the FindAllMarkers function with the Wilcox rank-sum algorithm and the following criteria: lnFC > 0.25, $p < 0.05$, min.pct > 0.1. Furthermore, in order to increase the accuracy, False discover rate (FDR) was calculated and significant genes (FDR<0.05) was selected for cell type annotation. To further identify cell types, clusters of a single cell type were selected for re-tSNE analysis. The fastMNN[3] function from the R package scran (version 1.10.2) was used to correct for the batch effect among samples by the mutual nearest-neighbor method.

CNV estimation

Endothelial cells, fibroblasts, and alveolar macrophages were used as references to

identify somatic copy number variations using the R package infercnv (version 0.8.2). We scored each cell for the extent of CNV signal, defined as the mean of the square of CNV values across the genome. Putative malignant cells were defined as those with a CNV signal of > 0.05 and a CNV correlation of > 0.5.

Differentially abundant subpopulation analysis

To detect differentially abundant subpopulations between the EGFR-positive and EGFR-negative groups, we used the DA-seq algorithm[4] calculated significant group specific cell, corrected by FDR, and labeled EGFR positive or negative specific regions using a contour plot.

Pseudo-time analysis

We conducted a single-cell trajectories analysis using Monocle2 (http://cole-trapnell-lab.github.io/monocle-release) with DDR-Tree and the default parameters. We first selected marker genes from the Seurat clustering result and raw expression counts of cell-passed filtering. Based on the pseudo-time analysis, “branch expression analysis modeling” (BEAM) was performed to determine branch fate. Pseudo-time BEAM genes were also calculated with p-value and q-value to adjust the gene expression significance. We selected the significant genes (q-val<0.0001) for further analysis.

Differential gene expression analysis

To identify differentially expressed genes, the FindMarkers function in Seurat Package was used with the Wilcox rank-sum algorithm using the following criteria: lnFC > 0.25,

$p < 0.05$, min.pct > 0.1. The p-value calculated by wilcox ransum test was adjusted by FDR.

QuSAGE gene enrichment analysis

To characterize the relative activation of a functional gene set in terms of GO terms, pathways, and hallmarks, or to identify the cell type/state, such as CAF types, macrophage types , we performed QuSAGE [5] (version 2.16.1) analysis.

Cell communication analysis

To evaluate cell–cell communication molecules, we performed a cell communication analysis based on CellPhoneDB [6], a public repository of ligands, receptors, and their interactions. Membrane, secreted, and peripheral proteins in clusters at various time points were annotated. We performed pairwise comparisons between all cell types. First, we randomly permuted the cluster labels of all cells (1,000 times as a default) and determined the mean of the average receptor expression level in a cluster and the average ligand expression level in the interacting cluster. For each receptor–ligand pair in each pairwise comparison between two cell types, this generated a null distribution. By calculating the proportion of the means which were equal or higher than the actual mean, we obtained a p-value for the likelihood of cell-type specificity of a given receptor–ligand complex (Significant relation method from https://www.cellphonedb.org/documentation). We selected the significant (p-value<0.05) cell-cell interaction relations for display.

Integration of multiple datasets

A lung carcinoma dataset (GSE146100) was analyzed in terms of the proportions of single cells and function alterations after immune therapy. The Seurat package (version: 2.3.4, https://satijalab.org/seurat/) was used for cell normalization and regression based on an expression table, according to the UMI counts of each sample and percent of mitochondria, to obtain scaled data. PCA of the scaled data was performed on all highly variable genes. The top 10 components were used for tSNE construction. The fastMNN function (k = 5, d = 50, approximate = TRUE, auto.order = TRUE) of the R package scran (v1.10.2) was used to apply the mutual nearest neighbor method, to correct for the batch effect among samples. Utilizing a graph-based cluster method, we acquired unsupervised cell cluster results based the top 10 principal components. The robust principal component analysis (RPCA) algorithm was applied for integration of scRNA-seq data (AnchorK = 5; AnchorFilterK = 200; AnchorKscore =30; AnchorDims =30; CellNumCutOff =30; K weight =100); the graph-based cluster method (resolution = 0.8) was used for clustering by cell type (epithelial cells, macrophages, CD8 T cells, CD4 T cells, stromal cells, and DCs).

Multiplex immunohistochemistry/ immunofluorescence assays

To confirm the status of T cells, macrophages, DCs, and immune checkpoint expression, formalin-fixed paraffin-embedded (FFPE) tumor tissue sections obtained from our patients were subjected to mIHC/IF. We analyzed the relative protein expression in four 5-um-thick slides from each patient. Antibodies were divided into four panels, as

follows. Panel 1: anti-CD3 (rabbit, 1:700, ab16669; Abcam, Cambridge, UK), anti-CD4 (mouse, 1:800, ZM0418; Zhongshan Jingqiao, Beijing, China), anti-CD8 (mouse, 1:600, CST70306; Cell Signaling Technology [CST], Danvers, MA, USA), anti-CD103 (rabbit, 1:800, ab224202; Abcam), anti-CXCL13 (rabbit, 1:2000, ab246518; Abcam); Panel 2: anti-CD3 (rabbit, 1:500, ab16669; Abcam), anti-CD8 (mouse, 1:400, CST70306; CST), anti-TIM3 (rabbit, 1:200, CST45208; CST), anti-LAG3 (rabbit, 1:200, CST15372; CST), anti-CTLA4 (rabbit, 1:1000, ab237712; Abcam); Panel 3: anti-CD3 (rabbit, 1:400, ab16669; Abcam), anti-CD8 (mouse, 1:300, CST70306; CST), anti-CD47 (rabbit, 1:5,000, ab226837; Abcam), anti-TIGIT (rabbit, 1:200, ab243903; Abcam), anti-PD1 (mouse, 1:50, CST43248; CST); and Panel 4: anti-CD68 (rabbit, 1:1000, BX50031; BioLynx, Ontario, Canada), anti-CD1c (mouse, 1:600, ab156708; Abcam), anti-CD11c (rabbit, 1:1000, CST45581; CST), anti-PD-L1 (rabbit, 1:400, CST13684; CST), and anti-Pan-CK (mouse, 1:200, CST4545S; CST).

We used the PANO7-plexIHC kit (#0004100100; Panovue, Beijing, China) for multiplex immunofluorescence staining. After sequential application of different primary antibodies, horseradish peroxidase-conjugated secondary antibody incubation and tyramide signal amplification using TSA Fluorescence Kits (Panovue) were performed. The slides were microwaved after each TSA operation. After all human antigens had been labeled, nuclei were stained with 4′-6′-diamidino-2-phenylindole (DAPI, Sigma-Aldrich, St. Louis, MO, USA). Stained slides were scanned using the Mantra instrument (PerkinElmer, Waltham, MA, USA), and fluorescence spectra were captured in the range of 420–720 nm with 20 nm interval. Individual scanned images

were combined to construct the image stack. The autofluorescence spectra of the tissue and each fluorescein were extracted from images of unstained and single-stained sections, respectively. The extracted images were used to establish a spectral library for multispectral unmixing via inForm image analysis software (PerkinElmer). Using this spectral library, we obtained reconstructed images of slices after removing autofluorescence.

Reference

1. Chen S, Zhou Y, Chen Y*, et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*. 2018; 34: i884-i90.

2. Dobin A, Davis CA, Schlesinger F*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*. 2013; 29: 15-21.

3. Haghverdi L, Lun ATL, Morgan MD*, et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*. 2018; 36: 421-7.

4. Zhao J, Jaffe A, Li H*, et al.* Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proceedings of the National Academy of Sciences of the United States of America*. 2021; 118: e2100293118.

5. Yaari G, Bolen CR, Thakar J*, et al.* Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic acids research*. 2013; 41: e170.

6. Vento-Tormo R, Efremova M, Botting RA*, et al.* Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature*. 2018; 563: 347-53.