Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing

Zhoufeng Wang
Sichuan University
Zhe Li
Singlera Genomics
Kun Zhou
West China Hospital, Sichuan University, Chengdu, Sichuan, China
Li Zhang
West China Hospital, Sichuan University, Chengdu, Sichuan, China
Ying Yang
Precision Medicine Research Center, West China Hospital of Sichuan University
Wenxin Luo
Department of Respiratory and Critical Care Medicine, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan
Wenliang Qiao
Lung Cancer Center, West China Hospital Sichuan University, Chengdu, Sichuan
Chengdi Wang
West China Hospital of Sichuan University
Gang Wang
West China Hospital, Sichuan University, Chengdu, Sichuan, China
Yinyun Ni
Precision Medicine Research Center, West China Hospital, Sichuan University
Shuiping Dai
Sichuan University, Chengdu, Sichuan, China
Tingting Guo
Department of Hematology, West China Hospital, Sichuan University, China
Guiyi Ji
West China Hospital, Sichuan University, Chengdu, Sichuan, China
Minjie Xu
Singlera Genomics
Yiying Liu
Singlera Genomics

Zhixi Su
Singlera Genomics
Guowei Che
Department of Thoracic Surgery, West China Hospital, Sichuan University
Weimin Li (weimi003@scu.edu.cn)
West China Hospital, Sichuan University, Chengdu, Sichuan

Article

Keywords: Lung adenocarcinomas (LUAD), single-cell RNA sequencing

Posted Date: January 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-127270/v1

Version of Record: A version of this preprint was published at Nature Communications on November 11th, 2021. See the published version at https://doi.org/10.1038/s41467-021-26770-2.

Abstract

Lung adenocarcinomas (LUAD) start as precancerous lesions such as atypical adenomatous hyperplasia (AAH), develop stepwise into adenocarcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA), then eventually progress toward invasive adenocarcinoma (IA). To date the cellular heterogeneity across these distinct clinical stages and the underlying molecular events driving tumor progression remain largely unclear. In this study, we performed single-cell RNA sequencing on 52 specimens from 25 patients spanning the four clinical stages. By assessing the expression pattern of marker genes among 268,471 cells, we identified 16 major cell types. We demonstrated that AT2 feature cell types (AT2-like cells) were associated with malignant composition. AT2-like subcluster emerged first in AAH and partially lost AT2 cell transcriptional identity, accompanied with a gain of stemness during cell transition. In addition, genes related to energy metabolism, ribosome synthesis were upregulated in the early stage of LUAD, leading us to identify new markers including miRNA10 and β-hydroxybutyric acid to diagnose early-stage LUAD noninvasively in the blood. We also identified MDK and TIMP1 as potential biomarkers to facilitate our understanding of LUAD pathogenesis. Taken together, our data identified a new mechanism in LUAD evolution, and provided a robust basis for diagnosis and treatment of LUAD.

Introduction

Lung cancer is the leading cause of cancer-related death worldwide, with adenocarcinoma representing the most prevalent subtype [1]. Among all lung adenocarcinomas (LUAD), the most prevalent subset is believed to develop via tumorigenesis and progress from atypical adenomatous hyperplasia (AAH) to adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and finally, overt invasive lung adenocarcinoma (IA) [2]. Over the past decade, efforts geared toward understanding LUAD progression have largely focused on profiling cancer cells with genetic aberrations [3]. Nonetheless, only increases in the frequency of genetic alterations during the progression to malignancy have been reported [4, 5], but the cellular heterogeneity and molecular events underlying cancer initiation and progression remain unclear.

The lung is composed of multiple cell types and fairly heterogeneous in terms of cellular composition. Basal, club, and ciliated cells are predominant in the proximal airway. Alveolar type 1 (AT1) and type 2 (AT2) cells are found in the alveolar wall [6]. Data from studies in mice suggested that AT2 cells, bronchioalveolar stem cells (BASCs), and club cells could be the cells of origin for LUAD [7]. Some of these mouse studies have been utilized in lung cancer lineage tracing; however, this could not be performed in human [8].

The advance in single-cell sequencing technologies have given rise to a comprehensive and unbiased analyses of cellular diversity within lung tissue [9, 10]. Single-cell RNA sequencing (scRNA-seq) has been successfully applied to identify unique cell types and understand the complex subpopulations in lung cancer. Previous scRNA-seq studies of lung cancer have been examined mostly stromal cells in limited LUAD subtypes [11], or charactering the molecular and cellular reprogramming of metastatic LUAD [12]. The

diversity of the cell states at different points and what roles the cells play in driving tumor progression are largely unknown.

In the present study, we performed scRNA-seq to survey 52 specimens including AAH, AIS, MIA, IA and non-malignant lung tissues. This results in, for the first time, a single-cell transcriptome atlas for all major subtypes of LUAD. Further analysis dissected the cellular and molecular characteristics of LUAD cells across different clinicopathological stages. We observed AT2 cells dedifferentiated into a stem-like state, a process to initiate and maintain tumor progression. We also showed that stem-like genes may be potential markers for diagnosing novel LUAD in the early stages. This scRNA-seq dataset unveils cellular dynamics and molecular features, which will allow us to decipher how the LUAD subtype progresses and evolves.

Results

Characterization of the cellular heterogeneity across four LUAD subtypes

A total of 52 freshly resected lung specimens of four subtypes (3 AAH, 5 AIS, 9 MIA, and 17 IA) were collected from 25 patients (**Fig. 1A** and **Table 1**), along with 18 adjacent non-malignant (also as normal) lung tissues from a distal region within the same lobe, which served as controls. Eight of the 25 patients presented multiple nodules (**Fig. S1**). For each specimen, we rapidly digested the freshly collected tissues to generate a single-cell suspension, isolated cells without enrichments for specific cell types, and generated scRNA-seq data with the 10x Chromium platform. We characterized the transcriptome of 140,556 cells from LUAD samples (P1-P22) at single-cell resolution by the V2 kits and validated our results with a separate dataset of 127,923 single cells from multiple nodules resected from an additional 3 patients (P23-P25) using the V3 kits (**Fig. 1A** and **Table 2**).

To distinguish cell populations based on distinct transcriptional profiles, we performed dimensionality reduction and unsupervised cell clustering using the Seurat package (version 3.0.3.9028). We identified 16 distinct clusters based on key marker genes that were further assigned into 16 major cell types (**Fig. 1B-1E**), including epithelial cells (ciliated, club, basal, AT1, AT2 and AT2-like cells) and stromal cells (endothelial cells, fibroblast cells, lymphocytes and myeloid cells). We profiled the transcriptomic characteristics of LUAD through comparative analysis between non-malignant and malignant cells from surgically resected specimens of four stages (**Fig. 1C-1G**). Most non-malignant cells were immune cells, and each cell cluster comprises cells derived from multiple different patients. The proportions of T lymphocytes and myeloid cells were highly reproducible across patients. Moreover, we confirmed the enrichment of B and T lymphocyte and the decline in natural killer (NK) and myeloid cells during tumor progression (**Fig. 1G** and **Fig. S2A-S2B**), indicating the activation of adaptive immune responses. In tumor tissues, we discovered that AT2 feature cell types (AT2-like cells) were associated with malignant composition, and were present at the onset of LUAD development (**Fig.1B and 1D**). The AT2-like cells were highly patient-specific and exhibited increasing cellular phenotypic heterogeneity with tumor progression (**Fig. 1D and 1F**). Notably, a very small percentage of cells expressing cell proliferation

markers was observed in our data, so we did not correct for the cell cycle effect (**Fig. S3A and S3B**). These results illustrated a wide range of intratumoral heterogeneity in LUAD transcriptome, which could be shaped by surrounding microenvironment during progression.

Next, we reanalyzed the published data of 2 LUAD patients (out of the total 8 patients) from the study by Lambrechts et al [11] by integrating our datasets, based on the same normalization and filtering parameters (**Fig. S4A-S4E**), and identified the same 16 cell clusters. Both datasets revealed highly consistent assignment of identified cell types. However, our dataset consisted of multiple instances of rare cell populations/clusters (such as ciliated cells, lymphocytes, and AT2-like cell subgroups) that were not present in the Lambrechts et al dataset. For example, the AT2-like cluster was enriched in premalignant AAH and AIS tumors (**Fig. S4A and S4B**). A validation cohort of an additional 127,923 single cells from 3 LUAD patients using the V3 kit (**Table 1**) revealed that epithelial and stromal cells could be assigned to subclusters, representing 90% of cell types in the original set of 22 patients (**Fig. S4C** and **S4D**). With this more comprehensive dataset, we further characterized the stromal cell and epithelial cell populations in a greater detail to better assess cell heterogeneity during LUAD progression.

Characterization of stromal cells in coordinating tumor microenvironment through tumor progression

The stromal cells are associated with tumors, could provide deeper insights into lung cancer biology [11]. To investigate stromal cell dynamics in the tumor microenvironment (TME), we analyzed the single-cell transcriptomes of endothelial cells (ECs), fibroblasts, lymphocytes, and myeloid cells from normal tissues and tumor tissues representing the four stages, as shown in **Fig.1B**. We detected 1, 925 ECs and five clusters based on marker genes; (**Fig. 2A-2D** and **Fig. S5A-5B**) these clusters included tip-like cells, tumor ECs, stalk-like cells, endothelial progenitor cells (EPCs) and lymphatic ECs. Most of the EC clusters observed belonged to normal tissues, and were assigned to known vascular cell types. For example, lymphatic ECs were enriched in normal tissue. Tip-like cells, stalk-like cells and EPCs were expressed in early-stage (AAH and AIS stage) tumors and normal tissues (**Fig. S5A**). Tumor ECs were identified in early-stage tumors, and demonstrated strong expression of PLVAP, GSN, and TSC22D1, which regulate the development and determine the cell fate of ECs. To learn more about the biology underlying these cell states, we used GSEA to compare the expression patterns in tumor ECs versus normal ECs (**Fig. 2D**). The results revealed Myc targets as the top enriched signature in tumor ECs. Indeed, earlier studies indicated that c-Myc is essential for tumor angiogenesis, glycolysis and oxidative phosphorylation, all of which promote vessel sprouting [13]. The most significantly downregulated pathway was involved in inflammatory responses, such as the INF-α and INF-γ responses. The endothelium represents the primary interface between circulating immune cells and the tumor, and plays important roles in relaying signals and presenting epitopes from the tissues it vascularizes to the immune system [14].

Fibroblasts have long been suggested to represent a heterogeneous population, but the extent of heterogeneity has remained unclear in LUAD progression [15]. Our subclustering of fibroblasts revealed six distinct clusters, including fibroblast-like cells, normal fibroblasts, smooth muscle cells, lipofibroblasts and myofibroblasts (**Fig. 2E-2H** and **Fig. S5A-5B**). The fibroblast-like cell (CFD and APOD) and

myofibroblast (ACTA2 and RGS5) expression patterns were reproducibly detected in AAH and AIS tumors, and could represent common features of the LUAD TME (**Fig. S5A**). Many studies have identified fibroblast-like cells and α-SMA-positive (ACTA2, smooth muscle α actin) myofibroblasts as cancer-associated fibroblasts (CAFs), promoting extensive tissue angiogenesis [16], and tumor progression[17]. Smooth muscle cells, comprising of the main fibroblast type in the vasculature and being linked to wound healing and angiogenesis [18], were observed in IA stage tumors and a few normal tissues. Further GSEA analysis of fibroblasts from normal and tumor tissues was performed (**Fig. 2H**). Cancer-derived fibroblasts are associated with epithelial-mesenchymal transition (EMT) and the immune response, such as showing strong IFN-γ and IFN-α responses. Therefore, cellular dynamics in ECs and fibroblasts supported a consistent phenotypic shift of stromal cells towards tissue remodeling and angiogenesis in LUAD.

Lymphocytes often play important roles in inflammation, cancer immune evasion, and responses to immunotherapy treatment [19]. In our dataset, subclustering of 61,196 detected lymphocytes revealed 10 clusters, annotated mainly as T cells, B cells and NK cells, with some other immune cell types (**Fig. 2I-2K** and **Fig. S5A-5C**). Consistent with previous findings, T cell-mediated cytotoxicity is critical for tumor cell clearance [20]. We observed that $CD8^+$ T and regulatory T (Treg) cells were enriched in the tumor, but $CD4^+$ T cells and natural killer cells were depleted during tumor progression (**Fig. S5A-5C**). Furthermore, Treg cells persist in the IA stage, exerted a suppressive mechanism of antitumor immunity during tumor progression. $CD8^+$ T cells and Treg cells were enriched in the tumor population, which was also reported in Lambrechts et al. [11]. B cells and plasma cells are rare in most samples. Using definitive tumor lymphocytes and innate lymphocytes, we constructed a transcriptional trajectory for the exhaustion of lymphocytes and found the key gene expression programs during tumor progression (**Fig. 2K**). Indeed, transcriptional states in the trajectory suggested that lymphocytes were not associated with changes in the tumor biology. Most of cell types were observed to be distinctly positioned at the branches. However, $CD8^+$ T and cancer cells were located in separate trajectory branches from the same ancestor, indicating their interactive differentiation states. Altogether, the changes in the cellular composition and gene expression phenotype of lymphocyte cells confirmed the direction of tumor immunity towards immune suppression in LUAD progression.

Myeloid cells play a critical role in maintaining tissue homeostasis, and regulate inflammation in the lung [21]. We dissected the gene signatures of the 8 myeloid clusters revealed in this study (**Fig. 2L**) including granulocytes, macrophages, NK cells and dendritic cells. Neutrophils were not recovered in our experimental procedures. Two macrophage types are known to populate the lung: the alveolar macrophage (AM) type, which highly expresses the MARCO, FABP4, and MCEMP1 genes; and the tumor macrophage (TM) type, which comprises the remaining tumor-enriched clusters. The AM type macrophages were mainly detected in normal and early-stage LUAD (**Fig. 2L-2N** and **Fig. S5A-5C**). By contrast, the TM transcriptional phenotypes were mostly present in late clinical stage IA tumors, and TMs shared high expression of SPP1, APOE, CCL2 genes, involved in apolipoprotein metabolism [22]. Trajectory analysis on TMs demonstrated a dynamic functional spectrum from AMs with regard to the

transcriptional trajectory (**Fig. 2N**). Dendritic cells mostly comprised myeloid cells, three DC subsets, DCs (CCL17), activated DCs (BIRC3, CCL22) and $CD141^+$ DCs (CPVL, KLRB1). $CD141^+$DCs mediated by NK cells, plays a key role in an inflammatory environment (**Fig. 2M**). On the other hand, cancer cells in myeloid cell clusters were mostly observed in late clinical stage IA tumors, which indicated that tumors communicating with immune cells after stage IA are capable of immune escape.

TME is heterogeneous and includes reprogrammed or activated immune cells, fibroblasts, and ECs. To characterize stromal cell heterogeneity in LUAD progression, we performed simultaneous immunofluorescence staining for ECs (CD31), fibroblasts (fibronectin) and immune cells (CD45) in normal tissues and tumor tissues from different stages (**Fig. S5D** and **Table 3**). An increase in the $CD45^+$ population was observed as the tumor progressed, suggesting that immune cells contribute to LUAD growth. On the other hand, we also examined the gene expression of subclusters in LUAD and patient survival (**Fig. S6**). Correlations among seven cell subtypes, such as CD4 T cells and T follicular helper cells, may reflect differences underlying the histopathology. These findings suggest that stromal cells and immune cells in the lung TME harbor both tumor-promoting and tumor-suppressing activities, which may predict clinical outcome.

Characterization of epithelial cell lineages across different stages

Lung epithelial cells have been studied extensively due to their medically important role in lung cancer and various pulmonary diseases such as asthma and fibrosis [21]. Here, we explored the intrinsic transcriptome of epithelial cells through comparative analysis between normal epithelium and tumor cells from different stages. We identified 15,984 epithelial cells and re-clustered them into 10 subclusters. (**Fig. 3A-3C**). Based on the expression of known markers, we found that the atlas mainly comprised of epithelial cells such as AT1 cells (AGER), AT2 cells (SFTPC), club cells (SCGB1A1), basal cells (Krt5), ciliated cells (FOXJ1) and AT2-like cells. As expected, normal epithelial cells mainly have five distinct subpopulations, including AT1, AT2 cells, club cells, basal cells and ciliated cells, expressing well-defined epithelial markers. Interestingly, the SFTPC gene (an AT2 cell marker) was highly expressed in non-malignant tissues and tumors at an early stage, but not in late-stage tumor tissues (**Fig. 3B**).

In tumor tissues, epithelial cells may contain residual non-malignant cells along with malignant tumor cells. To separate true tumor cells from a potential non-malignant population, we used inferCNV on scRNA-seq to identifying the malignant cells (**Fig. 3D-3E**). We used the cell types from adjacent lung tissues as a healthy reference to estimate the copy number variation (CNV) at different stages. Chromosomal amplification (red) and deletion (blue) were mapped to each chromosomal position (columns) across the single cells in **Fig. 3E**. We identified large-scale chromosomal CNVs in IA stage AT2-like cells, but did not observe any CNVs in other cell types.

To confirm the expression of identified epithelial cell marker genes, we performed immunofluorescent staining of different cell types to investigate the abundance and spatial localization of AT1 cells (AGER), ciliated cells (FOXJ1), club cells (SCGB1A1), basal cells (Krt5) and AT2 cells (SFTPC) (**Fig. S7A** and **Table**

3). Non-malignant human lung alveolar and bronchi tissues were double-stained with AGER and SFTPC antibodies. AT1 and AT2 cells were found to be preferentially localized in peripheral alveoli, while club cells, ciliated cells and basal cells were mainly distributed on the bronchial surface; these observations were consistent with a previous study [11]. In LUAD tissues, expression of the SFTPC gene was observed in MIA tumors, but less than 45% of the AT2 cells expressed SFTPC compared to that of AAH (**Fig. 3F**). AT1, basal cells, ciliated cells and club cells were not detected in MIA. In IA tumors, SFTPC gene expression was almost undetectable, and the alveolar structure was not recognizable. Therefore, SFTPC gene expression was significantly decreased during histopathological progression. These changes were associated with the previously described loss of expression of the lung lineage-defining transcription factor Nkx2-1 as well as loss of the AT2 markers SFTPC [23]. The results demonstrated that AT2 cells are the origin of adenocarcinomas and are defined as AT2-like cancer cells.

To further elucidate the differentiation process in LUAD, we constructed a lentiviral vector expressing EGFRVIII and Cre driven by the carbonic anhydrase II (CAII) promoter, which is expressed mainly in AT2 alveolar cells and was previously demonstrated to efficiently drive LUAD formation in mouse lungs [24]. After transfection of lentiviral vectors into mice, non-malignant and tumor lung tissues were resected after 3 and 7 months, and immunofluorescence stained (**Fig. S7B**). We found that the SFTPC was evenly distributed in the alveoli of non-malignant lungs. SFTPC gene expression was decreased in tumor tissue at the 3rd month and unevenly distributed and expressed at very low levels in the 7th month (**Fig. S7C**). Therefore, the results further supported that AT2 cells give rise to LUAD in spontaneous EGFRVIII mice and in human.

Transcriptional trajectory analysis of AT2 cells

To identify the key molecular events governing the cell-fate transition during progression from normal AT2 cells to cancer cells, we selected cell clusters that closely resemble those of AT2 and AT2-like cancer cells, and then tracked the gene expression changes along the trajectory from non-malignant tissues to AAH, AIS, MIA and eventually IA. We performed a pseudotime analysis with Monocle2 and observed non-random expression patterns (**Fig. 4A-4C**). The transcriptional states in the trajectory revealed normal differentiation paths as well as progression-associated changes in tumors. Non-malignant cells and cells at early clinical stages (AAH to AIS) gathered on one end, while cells from late-stage tumor tissues (MIA or IA) tended to be on the other end (**Fig. 4A**).

We identified 283 differentially expressed genes that exhibited dynamic expression over the pseudotime axis (q value < 0.05) and classified them into 4 groups (group 1 to group 4). Then, we ordered these genes along the temporal pseudotime and reconstructed a diffusion map (**Fig. 4B**). The expression profile of group 1 showed that the self-renewing AT2 genes were relatively quiescent (with high level of WIF1 inhibiting WNT) [25, 26] and had a high percentage of stem-like cell transcription and differentiation genes (LAMP3 and MUC1) expressed [9, 27]. By contrast, group 2 genes showed an expression pattern similar to the start of the dedifferentiation process, where lineage-specific genes such as negative regulation of the cell death response to stimulus (AQP5) [28], RNA biogenesis processing (RPL family) [29, 30], and

mitochondrial factors (MT-ND4 and MT-ND2) showed increased expression [31]. The expression pattern of group 3 genes was related to inflammatory response cytokine cell activation (FOSB, NFKB1), with visible EMT-related gene (vimentin) [32]. Lastly, the genes in group 4 were involved in extracellular matrix organization (TIMP1) [33], cell-cell signaling and regulation of cell migration (S100A4, VEGFR) [34]. Moreover, we also identified genes (MDK, SOX4 and LYZ) previously described during cancer evolution[1, 35]. Although each of these expression patterns emerged at a different specific time, they all persisted in tumors once they arose, such that more advanced tumors contained a greater assortment of cells with a higher diversity of states.

We then filtered out the genes expressed in fewer than 10% of total cells and performed gene regulatory network analysis on normal and AT2-like cells in order to examine the genetic interactions between the remaining 3,613 genes in the AT2 and AT2-like cell clusters (**Fig. 4C**). We found that genes upregulated in normal tissues are essential for lung function and homeostasis. In contrast, the genes upregulated in cancer are involved in metabolism, ribosomal activity, or MHC class II molecule expression, which suggests that these activities are essential during tumor progression. Notably, significantly downregulated genes in tumor ECs were related to immune activation, supporting a previous finding that tumor ECs suppress immune responses [11].

Loss of AT2 features and gain of stemness associate with cell transition

Multiple mouse studies have suggested stemness of AT2 cells, and the maintenance of this self-renewal activity is of great importance for cancer progression [36, 37]. We focused on clusters closing to normal AT2 cells, distinct subcluster that emerged first in AAH lost some AT2 cell transcriptional identity, but retained features of the lung epithelial lineage. Then we compared with the expression of normal AT2 cells, observed AT2-like cell subset with a signature of stemness genes, which were present in tumors throughout LUAD progression (**Fig. 4B-4C**). The result suggested that LUAD evolution is characterized by a loss of the AT2 feature of the lung lineage and the emergence of an alternative dedifferentiated stem-like state. These results were also consistent with previous report that suggested dedifferentiation of committed epithelial cells into stem cells in multiple diseases, especially cancer[38].

Epithelial cells transition to a mesenchymal state in tumor progression has been proposed [32]. Transcriptionally, EMT is defined by downregulation of the epithelial marker E-cadherin and upregulation of the mesenchymal transcription factors vimentin. To investigate the prevalence of EMT in LUAD progression, we examined the expression level of vimentin, E-cadherin and FOXM1 by RNA fluorescence in situ hybridization (FISH) (**Fig. 4E**). Expression of the E-cadherin decreased in conjunction with increased vimentin. We also found higher protein expression of FOXM1, which is a pro-stemness transcription factor associated with tumor proliferation in kidney and ovarian cancers [39, 40]. Our results suggest that FOXM1 could be a driver of dedifferentiation and proliferation in LUAD. These findings were further supported by tissue immunofluorescence and bioinformatics analysis of known epithelial and fibroblast markers. We next used indirect immunofluorescence staining to validate our scRNA-seq

findings at the protein level. Ki67 is a protein that expresses in all phases of the cell cycle except for quiescent cells in G_0. Interestingly, the staining intensity in our study showed high Ki67 expression (gray) in the early stages (AAH and AIS) and gradual decreases in expression as the tumor progressed (**Fig. S8**). This finding is of clinical importance, as the differences in Ki67 expression during LUAD progression indicate the cutoff values used for treatment decisions, which is consistent with our bioinformatic results and previous report [41]. EMT (vimentin, red) and angiogenesis (VEGF, green) related genes also gradually increased with tumor progression, as indicated by immunostaining.

There is a plethora of role of Wnt in stem cell self-renewal or lineage specific differentiation in diverse tissues in vivo [25, 42]. Recently, Wnt signaling was found to be amplified by engaging the leucine-rich repeat-containing G-protein-coupled receptor (Lgr) Lgr5, which is a marker for stem cells in multiple epithelial tissues and can drive lung adenoma progression in mouse models [25]. Consistent with the known role of the WNT signaling pathway, Lgr5 was reported to activate two WNT mediators: GPX2 and OLFM4 (**Fig. 4B**), which were shown to be increased during tumor evolution [43]. RNA FISH of four stage LUAD revealed an increase in the fraction of Lgr5 compared to normal tissues (**Fig. S9**), suggesting that the increased tumorigenic potential correlates with an increased stem-like signature. Differential expression of another stem-like gene, AQP5 gene is a compartment of WNT-driven invasive gastric cancer [28], observed in our results (**Fig. 4B**). Moreover, the expression of the stem-like genes IFI27 and S100A4 increased as the tumors progressed. Therefore, we speculate that AT2 cells undergo dedifferentiation to generate a stem-like state to initiate and maintain tumor progression.

Cell-cell crosstalk visualizing potential specific interactions in LUAD

Recent studies have demonstrated involvement of the specific hematopoietic stromal cell lineage and tumor epithelial cells in a cell-type-specific crosstalk-regulated network [44]. We used CellPhoneDB to identify the expression of potential crosstalk signaling molecules based on ligand-receptor interactions [45]. In **Fig. 5A**, we showed that both normal and tumor epithelial cell clusters expressed genes found within myeloid cells, epithelial cells (especially AT2/AT2-like), and fibroblasts, which suggested a possible interaction between epithelial and stromal cells. **Fig. 5B** displays the detailed receptor-ligand interaction pairs (basically gene pairs) for the cell types of indicated cell types. We focused on the gene pairs in the cell types that exhibited a strong interaction, and extracted the rows with all gene pairs (p value <0.001). As a result, the number of interacting cells was dramatically elevated in AT2-like cells, communicating with immune cells. AT2 cells expressed higher levels of LGALS9, the receptors of COLEC12 and MRC2 in DCs, granulocytes and macrophages. LGALS9 on AT2 cells has previously been found to promote immune suppression via T-cell and macrophage inhibition [46]. By contrast, AT2-like cells express high levels of ANXA1, MDK, FN1 and CCR1. The expression pattern of the FN1–A4B1 (A4B7) ligand-receptor complex indicates the existence of functional interactions between AT2-like cells and immune cells. Here, we pinpointed that AT2 cell interacted specifically with the myeloid cell subset by LGALS9 receptor, but AT2-like cells tightly interacted with myeloid cells via ANXA1, FN1and MDK (**Fig. 4C**). The results suggested that these two types of ligand-receptor interactions could be another new immune checkpoint,

as potential novel immunotherapy target for LUAD. On the other hand, endothelial cells and fibroblasts prominently expressed ligands such as FN1, CCL2 and CCL12, also could match immune related receptors.

Independently, we performed flow cytometry to measure the distribution of the two populations and the preferential expression of immune cells in AT2-like cells (**Fig. 4D**). There was elevated expression of CD45 in AT2-like cells ($Epcam^{+}CD45^{+}$ cells) as LUAD progressed. These results were consistent with our scRNA-seq data showing higher levels of expression of PRF1, GNLY, GZMA and GZMB genes in this subset, thus confirming that immune cells communicating tumor cells towards immune suppression in LUAD.

Quantitative detection of key biomarkers in plasma of different stage LUAD

Biomarkers that can effectively diagnose lung cancer in the early stages or confirm the presence of metastasis can guide clinical intervention and treatment. Multiple research groups have recently turned to blood-based biomarkers to detect lung cancer [47, 48]. At present, accepted clinical blood tests for LUAD mainly based on carcinoembryonic antigen (CEA), but recent studies have shown that sensitivity for early-stage lung cancer is limited [49].In our dataset, genes related to energy metabolism and ribosome synthesis were upregulated in the early stage of LUAD, which may favor transcription and metabolism required for tumor survival and growth. Therefore, we performed quantitative detect of protein, miRNA and metabolite biomarkers associated with metabolism and ribosome synthesis genes for LUAD diagnosis. Among them, miRNA-10a has been speculated to control synthesis via stimulation of ribosomal protein mRNA translation and ribosome biogenesis [50]. On the other hand, β-hydroxybutyric acid are likely to be produced by nearby or adjacent fibroblasts to provide energy to tumor cells, and has been detected in lung cancer patients [48, 51]. Therefore, we carried out RT-PCR or ELSA to identify plasma miRNA-10a and β-hydroxybutyric acid levels to distinguish LUAD patients at different stages of disease and healthy controls (**Fig. 6A** and **Table 4**). While plasma CEA level showed no difference between specimens from healthy controls and patients with early-stage cancer (AIS and MIA), the levels of miRNA10 and β-hydroxybutyric acid were significantly higher in the plasma of the early-stage LUAD patients ($P<0.01$) (**Fig. 6A**). Additionally, miRNA10a levels were found to be significantly upregulated in all four clinical stages of NSCLC [52] ($P=0.002$). Based on this analysis, miRNA-10a and β-hydroxybutyric acid appeared to be appropriate plasma biomarkers for distinguishing patients with early-stage disease and healthy controls.

The changes in marker gene expression were also examined along pseudotime. Notably, TIMP1 and MDK (**Fig. 3B**) were highly expressed at the late clinical stage. Elevated levels of MDK, a product of lysine decarboxylation, were also identified as one of the most important features in discriminating stage III LUAD (**Fig. 6A**). Moreover, TIMP1 also stand out as a critical feature in stage III LUAD. Previously, MDK and TIMP1 have been reported to regulate metabolism in metastases by activating the PI3K/Akt pathway [33, 53] , which were also validated in our study by immunostaining (**Fig. 6B**). Taken together, these results highlighted the potential of new marker genes such as miRNA10 and β-hydroxybutyric acid to diagnose

early-stage LUAD, and suggested MDK and TIMP1 as potential biomarkers to facilitate our understanding of LUAD pathogenesis.

Discussion

The TME is composed of multiple cell types [11]. Such cellular complexity of tumors is accompanied by heterogeneity within each cell type [54]. Here we provide a high-resolution scRNA-seq data set of LUAD at four stages to construct a lineage map of LUAD progression. First, in terms of the catalog of stromal cells in TME, tumor ECs acquired highly angiogenic, yet immune compromised properties. Second, fibroblast-like cells and myofibroblasts are CAFs that promote tumor progression. Third, $CD8^+$ T cells and Tregs persisted in the IA stage, providing a suppressive mechanism of antitumor immunity during tumor progression. Fourth, TMs transcriptional phenotypes, which are involved in apolipoprotein metabolism, were observed mostly in late clinical stage IA. Mechanistically, we have shown that alteration in the immune response during tumor progression is associated with immune-suppression via immune escape. However, it is still largely unknown how regulators of the immune response act on tumor progression. Future research on this topic will inform us more on how tumors communicate with stromal cells and alter their cell status or identity.

The heterogeneity of tumor cells represents a major challenge in oncology. We found the transcriptional heterogeneity grew dramatically during tumor progression, but the process was stereotypical and reproducible across individual tumors. Different epithelial cell types exhibit unique molecular signatures. Early cancer cells closely resembled AT2 cells, which underwent dedifferentiation to a stem-like state and further initiate tumor progression, eventually giving rise to the heterogeneous populations observed in LUAD (**Fig. 7**). In the normal alveoli, AT2 cells self-renew under homeostatic conditions and can generate stem-like cells after injury or gene mutation [6]. A previous study also revealed that alveolar progenitor cells (such as AT2 cells are responsible for the repair of alveoli [55]. Our work showed that the expression of tissue-specific marker genes such as SFTPC and SFTPB was downregulated and that of stem cell signaling factors such as CD44 was upregulated. We observed stem-like population in early-stage tumor cells that likely continued proliferating with increased epithelial diversity. Eventually, late-stage tumor cells exhibited profound cellular and genetic heterogeneity. In addition, we found that MDK and TIMP1 are potential biomarkers to facilitate our understanding of LUAD pathogenesis.

Our results suggested that disrupting of normal developmental programs is a major organizing principle in the acquisition of an AT2-like state. We observed that $Lgr5^+$ LUAD cells display persistent proliferative potential, followed by several alternative differentiation programs mimicking the primordial lung and finally the emergence of cells with a mesenchymal phenotype, indicating complete EMT (**Fig. 7**). E-cadherin, IFI27 and S100A4 could induce EMT and promote cancer cell migration and stemness [34, 56, 57, 58]. Our study represents an important step to understand intra-tumoral expression in epithelial tumor progression.

Immune cells can communicate via ligand-receptor interactions [44], so targeting cell-cell interactions have become a useful tool in clinical practice. For example, the immune checkpoint inhibitor ipilimumab targets the binding of ligands to CD28 or CTLA4, and both pembrolizumab and nivolumab target the interaction between PD1 and PDL1 [59]. Malignant cells differentially induce the activation of cellular and humoral immune responses in individual patients; however, they concomitantly provide inhibitory signals to induce immune exhaustion. We observed both normal and cancerous epithelial cell clusters expressed myeloid cells genes, which suggested a possible interaction between epithelial and immune cells. AT2-like cells express high levels of ANXA1, MDK, FN1 and CCR1, then from ligand-receptor complex with immune cells, but AT2 cell interacted specifically with the myeloid cell subset by LGALS9 receptor. The results suggested that these two types of ligand-receptor interactions could be potential novel immunotherapy target for LUAD. On the other hand, cancer cells in T-cell clusters were mostly in late clinical stage IA, which indicated tumor communication with immune cells, inclined to promote immune escape in IA stage. Several studies have demonstrated that tumor-infiltrating $CD8^+$ T lymphocytes impair production of effector cytokines in advanced NSCLC [60]. Overall, this large compendium of single-cell data could help explain the tumor-immune interactions by providing insights into the composition, states and dynamics of T cells in LUAD progression. However, our knowledge of which interactions occur in tumors and how these interactions affect the prognosis are still limited. There is a need to elucidate the spectrum of cell-cell interactions occurring in the TME and how these interactions affect patient outcomes.

In the present study, ribosomal and mitochondrial genes were highly expressed in AT2-like cancer cells at an early clinical stage, which favored transcription and metabolism to promote tumor progression. It is possible that an increase in ribosomal and mitochondrial gene expression could be an early indicator of lung cancer. Therefore, exploiting these differences to develop sensitive and accurate strategies for the early detection of lung cancer are essential to improve lung cancer survival. Consistent with our results, several other recently published studies have detected the same trend in plasma β-hydroxybutyric acid from lung cancer patients. The biological functions of β-hydroxybutyric acid are quite diverse including energy metabolism, epigenetic regulation, and oxidative stress response [61]. miRNA10a may positively control global protein synthesis via stimulation of ribosomal protein mRNA translation and ribosome biogenesis. For example, increased expression of RPS11 and RPS20 is associated with a stress-resistant glioblastoma stem cell phenotype and poor prognosis [62]. Additionally, silencing RPL39 expression was found to impact breast stem cell abilities, such as self-renewal and metastatic potential [30]. Therefore, we concluded that the scRNA-seq may provide a high-resolution method for the identification of molecular markers to precisely detecting the presence of early neoplastic sites and the onset of rare neoplastic cells in the clinic. These data suggest that ribosomal and stem-like genes are potential targets for developing diagnostic strategies and therapeutic interventions for lung cancer.

Conclusions

In summary, we constructed a single-cell transcriptome atlas of premalignant lung lesions and LUAD of all major clinical stages, to find out the heterogeneity at different stages on the basis of the

transcriptome, and the transcriptional landscape of lung epithelial cells. Specifically, we found that each epithelial cell type exists in a unique molecular state, but the early cancer cells are highly similar to AT2 cells, which undergo dedifferentiation to generate a stem-like state to initiate and maintain tumor progression, and to give rise to the heterogeneous populations observed in LUAD. In addition, our research has shown that miRNA10a, β-hydroxybutyric acid, MDK, and TIMP1 are potential biomarkers to facilitate understanding of LUAD pathogenesis.

Materials And Methods

Patients and clinical information

This study was designed to characterize the cellular heterogeneity and molecular events underlying LUAD initiation and progression. We performed single-cell RNA sequencing on 52 specimens from 25 patients spanning the four clinical stages (AAH, AIS, MIA, IA), with 18 of matched non-malignant lung samples. Patients diagnosed with AAH, AIS, MIA, and IA according the 2015 WHO classification [63], were enrolled with informed consent from West China Hospital of Sichuan University, China. The clinical characteristics of these participants, including age, gender, pathology, and tumor stage were recorded at recruitment (Table 1). All patients received surgical treatment and none of them underwent neoadjuvant therapy before surgery. Tumors and matched distal normal lung tissues were obtained during surgery. Cancer clinical stage was defined according to the 8th edition of the American Joint Committee on Cancer (AJCC) TNM stage system.

Sample preparation

Resected tumors were transported in Hank's Balanced Salt Solution (HBSS, Life Technologies) on ice immediately after surgical procurement. The tumor sample was subsequently divided into two pieces, and a small fragment was stored in liquid nitrogen for tissue staining. The remainder of the tumor was minced with scalpels into tiny cubes <0.5 mm^3 and transferred into a 15mL conical tube (BD Falcon) containing 8mL pre-warmed HBSS, 1mg/mL collagenase I and 0.5mg/mL collagenase IV. Tumor pieces were digested on Tube Revolver (Thermo) for 30 minutes at 37°C. This suspension was then filtered using a 70μm nylon mesh (BD Biosciences) and residual cell clumps were discarded; then the cell pellet was resuspended in red blood cell lysis buffer. Following a 5min incubation at room temperature, samples were centrifuged to discard the supernatant and re-suspend the cell pellet in PBS with 0.04% FBS. Cell sorting was performed with a MoFloAstrios EQ (Beckman coulter). The live cells were used for single-cell experiments after the dead cells being eliminated by excluding 7-Aminoactinomycin D (Life Technologies) cells.

scRNA-seq Library Preparation and Sequencing

Single-cell suspensions were converted to barcoded scRNA-seq libraries by using the Chromium Single Cell 3’Library, Gel Bead & Multiplex Kit and Chip Kit (10x Genomics) following the manufacturer's instructions, aiming for an estimated 5,000 cells per library. The sequencing ready library was purified

with SPRIselect, quality controlled for sized distribution and yield (LabChip GX Perkin-Elmer) and quantified using qPCR (KAPA). Libraries were sequenced on an Illumina NovaSeq-6000, and mapped to the human genome (build hg19) using CellRanger (10x Genomics).

Single-cell RNA sequencing analysis and identification of marker genes

Raw gene expression matrices generated per sample using CellRanger (version 3.0.0) were combined in R (version 3.6.3) and converted to a Seurat object using the Seurat R package (version 3.0.3.9028). Cells that had either more than 20,000 UMIs, over 3,000 or below 300 expressed genes, or over 10% UMIs derived from mitochondrial genome were removed. After filtering, the gene expression matrices were normalized to total cellular read count, original sample identity, and mitochondrial read count using linear regression as implemented in Seurat's Regress Out function. As a result, none of the principle components subsequently identified were correlated with transcript count (data not shown). To reduce dimensionality of this dataset, the variably expressed genes were summarized by principle component analysis, with the first 100 principle components further summarized using UMAP dimensionality reduction using the default settings of the RunUMAP function. Clustering was conducted with the FindClusters function using 50 PCA components with resolution parameter set to 2. Cell clusters in the resulting two-dimensional representation were annotated to known biological cell types using canonical marker genes. Very few cells were positive for cell proliferation markers, so we did not correct for effects of cell cycle in the analysis.

To identify marker genes of cell clusters, we contrasted cells from one particular cluster to those in all other clusters using the Seurat FindAllMarkers function. Marker genes were required to have an average expression in one particular cluster to be >2.5-fold higher than that in the other clusters.

Gene set enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA) is a widely used approach to test if a particular gene set is enriched at the top of a ranked gene list [64]. The 'fgsea' package v1.8.0 was used with default settings together with annotated Hallmark gene sets from the 'msigdbr' package [65]. The top 50 pathways ranked by adjusted P value (padj < 0.005) were plotted in the visualization.

Trajectory analysis

In order to generate a trajectory, we generated randomly sampled subset of malignant cells from each clinicopathological stage among the epithelial cells in the lung tumor tissue samples. Next, we employed the Monocle2 algorithm using the gene-cell matrix in the scale of UMI counts extracted from Seurat subset as input, and new Cell Data Set function was called to create an object with the parameter negbinomial size as the expression Family [66]. The cell trajectory was inferred using default parameters after dimension reduction and cell ordering.

inferCNV and clonality analysis

For the inferCNV analysis the following parameters were used: 'denoise', default hidden markov model (HMM) settings, and a value of 0.1 for 'cutoff'To reduce the possibility of false positive, CNV calling of the default Bayesian latent mixture model was implemented to identify the posterior probabilities of alterations in each cell. Low-probability CNVs were filtered using the default value of '0.5' as threshold. To determine the clonal CNV changes in each tumor the "subcluster" method was utilized on the CNVs generated by the HMM. GRCh37(hg19) cytoband information was used to convert each CNV to a p- or q-arm level change for simplification based on its location. Each CNV was annotated to be either a gain or a loss. After data conversion, subclones containing identical arm level CNVs were collapsed. ChrX & Y and Mitochondrial CNVs were excluded from this analysis. For data visualization, the UPhyloplot2 (https://github.com/harbourlab/UPhyloplot2) is used for evolutionary trees with default parameters. A scalable vector graphics (.svg) file visualizing the phylogenetic tree was generated for each sample.

Gene regulatory network analysis

We constructed the gene regulatory network in normal lung cells and cancer cells using the bigSCale2 algorithm (https://github.com/iaconogi/bigSCale2) [67]. Briefly, the expression data of the total 13,461 cells in the AT2, AT2-like cell clusters was extracted using Seurat and then combined into a sparse expression data matrix. We eliminated the genes expressed in less than 10% of total cells, leaving us with 3,613 genes for the network analysis. The resulting matrix was then passed to bigSCale2 for the construction of the network under the "direct" clustering parameter and by retaining only the genetic interactions with the correlation coefficient > 0.75. The network was then visualized in Cytoscape 3.8.0 (https://cytoscape.org/) with the Prefuse Force Directed Layout.

Cell–cell interaction network analysis

We mapped the cell-cell interaction and receptor-ligand pair between all major cell types using CellPhoneDB (www.cellphonedb.org) [68, 69]. The potential interaction between the two cell types were inferred through gene expression levels through 1000 permutation tests. Then the resulting adjacency matrices were generated for all cell–cell interactions and visualized on heatmap. We applied the following filtering steps: cell–cell interactions within identical cellular lineages were excluded, and only visualized the gene pairs from receptor-ligand interactions of the cell types we are interested in, and satisfied with combined p-value of 0.001 (multiplying all p-values within each gene-pairs).

TCGA survival analysis

To assess the correlation of specific cell types to lung adenocarcinoma patient survival, we downloaded LUAD RNA-seq data (TCGA-LUAD) as well as clinical data using the Bioconductor TCGAbiolinks package (version 2.2.10). 515 LUAD RNA-seq data (TCGA-LUAD) as well as clinical data were used. For the cell cluster identified in the study, top 10 marker genes in each cluster were ranked by logFC, and averaged per patient and grouped into high and low ones by the median expression for the corresponding cluster. The survival analyses were performed using R package survival (version 3.2-3) and survminer (version 0.4.7).

Kaplan–Meier curve and Cox regression P values, obtained after correction for age, gender, and tumor stage, were used to compare survival difference between the high and low groups.

EGFR VIII mutate mouse model

The lentiviral vector pTomo-EGFRVIII-IRES-Cre was constructed based on the backbone of the pTomo vector [24]. Lentiviral DNA vectors were co-transfected into HEK293T cells with the packaging plasmids pCMVΔ8.9 and pMD2.G at a ratio of 10:5:2. The copy number of lentiviral particles was confirmed using quantitative RT-PCR with U5 primers (U5-F, 5'AGCTTGCCTTGAGTGCTTCA3' and U5-R, 5'TGACTAAAAGGGTCT-GAGGG3'). Mice were subjected to general anesthesia with 2.5% avertin by intraperitoneal injection (15mL/kg). The mice were placed with their head forward in a supine position and $2x10^5$ lentivirus particles in 70 μL PBS was gradually released into the nostrils with sterile pipette tips. Finally, the mice were allowed to slowly wake up, and tumorigenesis was subsequently observed. We used mice for each lineage time point (3 months and 7 months) for Immunohistochemistry and immunofluorescence staining. All animal protocols were conducted as approved by the West China Hospital Institutional Animal Care and Use Committee.

Immunofluorescence staining

Flow cytometry

Dissociated cells were multi-stained with two antibodies at 4 °C for 30min, and then washed once with phosphate buffered saline. All antibodies were used at concentrations recommended by the manufacturer. After being filtered through a round-bottom tube with a 40- simultaneously stained cells with anti-human EpCAM-PE (8995s, CST) and CD45-PE (555483, BD) antibodies, the cells were fixed in 2% paraformaldehyde/PBS, permeabilized in Intracellular staining Perm Wash Buffer (BioLegend, San Diego, CA, USA). Then cells were fixed in 2% paraformaldehyde/PBS, permeabilized in Intracellular staining Perm Wash Buffer (BioLegend, San Diego, CA, USA). For the sorting of immune-epithelial communication in the in different stages, cells were stained for EPCAM and CD45.

In situ RNA-FISH

Formalin-fixed paraffin-embedded biopsies were sectioned to generate 5μm-thick sections within a week prior to perform in situ hybridization. All materials, including the microtome and blade, were sprayed with RNase-away solution prior to use. Slides were baked for 1 hour in a 60ºC dry oven the night before, and stored overnight at room temperature in a slide box with a silicone desiccator packet, and with seams sealed using parafilm. Sections were deparaffinized and dried at room temperature. Protease activity was blocked and the slides were incubated with probes. The fluorochromes Cy3 (Lgr5) and DIG (E-cadherin, FOXM1, Vimentin and EPCAM) were bound to the probes. The nuclei were counterstained with DAPI. The slides were covered with a coverslip and imaged under the Nikon Imaging system (NIKON DS-U3) and Caseviewer software. Imaging parameters were kept consistent for all images within the same

experiment, and any post-imaging manipulations were performed equally on all images from a single experiment.

Quantitative real-time PCR

A total of 93 LUAD patients and 30 healthy controls with comparable age and gender profiles were recruited. Among the cancer patients, 25 patients were male and 68 were female. Patients (and controls) with a history of any liver or kidney disease, and any previous treatment with anti-neoplastic drugs were excluded from this cohort. Blood sample (5-8mL peripheral venous fasting blood) was obtained before surgery, and immediately centrifuged; the plasma was stored at -80°C. Total RNA was isolated from cultured cells with Trizol reagent (Invitrogen, USA). The GAPDH was used as an internal control for each specific gene. For miRNA isolation, 200 μL yielded plasma was used for miRNA extraction using isolated QIAGEN miRNeasy Serum/Plasma Kit, (Cat No. 217184) according to manufactures' guideline. Quantitative real-time PCR (qRT-PCR) was performed on the IQ5 Real-time PCR system (Bio-Rad, Hercules, CA, USA), 1μg used for each sample. The relative transcript levels of miRNA-10a were calculated using the 2-DCT method. U6-snRNA acted as an external control, and all primers were synthesized by TsingKe Co. (Chengdu, China). Three independent experiments were performed to analyze the relative gene expression and each sample was tested in triplicate.

ELISA

The CEA (Abnova, ABIN6730886), β-hydroxybutyric acid (Biovision, K651-100), MDK (Abnova, ABIN924788) and TIMP1 (Abnova, ABIN365420) concentrations in the plasma from LUAD patients and 30 healthy controls were detected following the manufacturer's instructions. Optical density was measured in each well using a microplate reader Epoch BIO-TEK Instruments (Pittsfield, MA., USA). The concentration of cytokines in the samples was determined by comparing the optical density values of the samples to the standard curve.

Statistics and reproducibility

No statistical method was used to predetermine sample sizes. Samples were processed for scRNA-seq (10x Genomics) soon after resection in the operating room. As a result, samples from different patients were processed in separate experiments. Survival probabilities were estimated using the Kaplan–Meier method and differences between Kaplan–Meier curves were compared by using the log-rank test. Univariable and multivariable Cox proportional hazard regression models were used to identify independent prognostic factors. Hazard ratios (HR) with their 95% confidence intervals (CI) and corresponding P values were presented. All statistical tests were two-sided, and a P value lower than 0.05 was considered as statistically significant. All analyses were performed with SPSS V18.0 software (SPSS Inc., Chicago, IL, USA).

Declarations

Acknowledgements: We are thankful to all the patients for their voluntary participation in the study. We thank Dr. Kun Zhang from the University of California San Diego, and Wan Xiong from West China Hospital for the valuable discussion. We thank Dr. Dong Yang from the laboratory of animal tumor models West China Hospital, for Providing EGFR VIII mutate mouse model.

Author' contributions: Z. Wang, W. Luo and W. Qiao performed the experiments; G. Wang and Y. Ni performed sequencing experiments and processed the data; Z. Li, M. Xu, Y. Liu, Z. Su performed bioinformatic analyses; K. Zhou, C Wang, G, Ji consented the patients and collected the samples; S. Dai assisted in participant selection, consent, clinical information and procurement of tissue; L Zhang performed histological evaluations; Z. Wang, and Y. Yang, T. Guo performed fresh tissue dissociations, protein immunostainings, microscopy and imaging; W. Li and G. Che provided clinical insights; Z. Wang and Z. Li analyzed and interpreted the data. Z. Wang and Z. Li conceived of the experiments and wrote the manuscript; All authors reviewed and edited the manuscript.

Ethics approval and consent to participate: This study was approved by the local ethics committee at the West China Hospital of Sichuan University (Ethics: project identification code: 2018.270) and we complied with all relevant ethical regulations.

Availability of data: All relevant data are available via the data center of precision medicine in West China Hospital, Sichuan, China (https://pms.cd120.com/wzf01/index.html).

Competing interests: The authors declare no competing interests.

Funding: This work was supported by National Natural Science Foundation of China (Nos. 81871890, 91859203 to W Li, and 81802300 to Z Wang); China Postdoctoral Fund (No. 2019T120850 to Z Wang), Sichuan Science and Technology Program (No. 2019YJ0159 to Z Wang).

References

1. Gridelli C, *et al.* Non-small-cell lung cancer. *Nat Rev Dis Primers* **1**, 15009 (2015).
2. Inamura K. Clinicopathological Characteristics and Mutations Driving Development of Early Lung Adenocarcinoma: Tumor Initiation and Progression. *Int J Mol Sci* **19**, (2018).
3. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
4. Izumchenko E, *et al.* Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. *Nat Commun* **6**, 8258 (2015).
5. Hu X, *et al.* Multi-region exome sequencing reveals genomic evolution from preneoplasia to lung adenocarcinoma. *Nat Commun* **10**, 2978 (2019).
6. Swanton C, Govindan R. Clinical Implications of Genomic Discoveries in Lung Cancer. *N Engl J Med* **374**, 1864–1873 (2016).

7. Cheung WK, Nguyen DX. Lineage factors and differentiation states in lung cancer progression. *Oncogene* **34**, 5771–5780 (2015).
8. Marjanovic ND, *et al.* Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**, 229–246.e213 (2020).
9. Treutlein B, *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-sEq. *Nature* **509**, 371–375 (2014).
10. Montoro DT, *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
11. Lambrechts D, *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med* **24**, 1277–1289 (2018).
12. Kim N, *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Communications* **11**, (2020).
13. Baudino TA. c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression. *Genes & Development* **16**, 2530–2543 (2002).
14. Kambayashi T, Laufer TM. Atypical MHC class II-expressing antigen-presenting cells: can anything replace a dendritic cell? *Nat Rev Immunol* **14**, 719–730 (2014).
15. Kalluri R. The biology and function of fibroblasts in cancer. *Nature Reviews Cancer* **16**, 582–598 (2016).
16. Vong S, Kalluri R. The Role of Stromal Myofibroblast and Extracellular Matrix in Tumor Angiogenesis. *Genes & Cancer* **2**, 1139–1145 (2011).
17. Hinz B, *et al.* Recent developments in myofibroblast biology: paradigms for connective tissue remodeling. *Am J Pathol* **180**, 1340–1355 (2012).
18. Avila MA, Rockey DC, Weymouth N, Shi Z. Smooth Muscle α Actin (Acta2) and Myofibroblast Function during Hepatic Wound Healing. *PLoS ONE* **8**, (2013).
19. Thommen DS, Schumacher TN. T Cell Dysfunction in Cancer. *Cancer Cell* **33**, 547–562 (2018).
20. Guo X, *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature Medicine* **24**, 978–985 (2018).
21. Vieira Braga FA, *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat Med* **25**, 1153–1163 (2019).
22. Zhang L, *et al.* Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. *Cell* **181**, 442–459 e429 (2020).
23. Snyder EL, *et al.* Nkx2-1 represses a latent gastric differentiation program in lung adenocarcinoma. *Mol Cell* **50**, 185–199 (2013).
24. Yang D, *et al.* HUWE1 controls the development of non-small cell lung cancer through down-regulation of p53. *Theranostics* **8**, 3517–3529 (2018).
25. Tammela T, *et al.* A Wnt-producing niche drives proliferative potential and progression in lung adenocarcinoma. *Nature* **545**, 355–359 (2017).

26. I Ramachandran VG, E Gillies, I Fonseca, SM Sureban, CW Houchen, A Reis, L Queimado. Wnt inhibitory factor 1 suppresses cancer stemness and induces cellular senescence. *Cell Death and Disease* **5**, (2014).
27. Schroeder JA, *et al.* MUC1 overexpression results in mammary gland tumorigenesis and prolonged alveolar differentiation. *Oncogene* **23**, 5739–5747 (2004).
28. Tan SH, *et al.* AQP5 enriches for stem cells and cancer origins in the distal stomach. *Nature* **578**, 437–443 (2020).
29. Amandine Bastide AD. The ribosome, (slow) beating heart of cancer (stem) cell. *Oncogenesis* **7**, (2018).
30. Dave B, *et al.* Targeting RPL39 and MLF2 reduces tumor initiation and metastasis in breast cancer by inhibiting nitric oxide synthase signaling. *Proc Natl Acad Sci U S A* **111**, 8838–8843 (2014).
31. Angelidis I, *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *bioRxiv*, (2018).
32. Dongre A, Weinberg RA. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat Rev Mol Cell Biol* **20**, 69–84 (2019).
33. Song G, *et al.* TIMP1 is a prognostic marker for the progression and metastasis of colon cancer through FAK-PI3K/AKT and MAPK pathway. *J Exp Clin Cancer Res* **35**, 148 (2016).
34. Hou S, *et al.* S100A4 promotes lung tumor development through beta-catenin pathway-mediated autophagy inhibition. *Cell Death Dis* **9**, 277 (2018).
35. Olmeda D, *et al.* Whole-body imaging of lymphovascular niches identifies pre-metastatic roles of midkine. *Nature* **546**, 676–680 (2017).
36. Ahmad N. Nabhan DGB, Pehr B. Harbury, Mark A. Krasnow, Tushar J. Desai. Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science* **359**, 1118–1123 (2018).
37. Desai TJ, Brownfield DG, Krasnow MA. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature* **507**, 190–194 (2014).
38. Tata PR, *et al.* Dedifferentiation of committed epithelial cells into stem cells in vivo. *Nature* **503**, 218–223 (2013).
39. Xue YJ, Long DZ, Zou XF, Wang XN, Zhang GX, Yuan YH, Wu GQ, Yang J, Wu YT, Xu H, Liu FL, Liu M. Overexpression of FoxM1 is associated with tumor progression in patients with clear cell renal cell carcinoma. *Journal of Translational Medicine* **10**, (2012).
40. Wen N, *et al.* Overexpression of FOXM1 predicts poor prognosis and promotes cancer cell proliferation, migration and invasion in epithelial ovarian cancer. *Journal of Translational Medicine* **12**, (2014).
41. Focke CM, Decker T, van Diest PJ. Intratumoral heterogeneity of Ki67 expression in early breast cancers exceeds variability between individual tumours. *Histopathology* **69**, 849–861 (2016).
42. Pongracz JE, Stockley RA. Wnt signalling in lung development and diseases. *Respir Res* **7**, 15 (2006).

43. Li H, *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* **49**, 708–718 (2017).
44. Kumar MP, *et al.* Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Rep* **25**, 1458–1468 e1454 (2018).
45. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols* **15**, 1484–1506 (2020).
46. Seifert AM, *et al.* Detection of pancreatic ductal adenocarcinoma with galectin-9 serum levels. *Oncogene* **39**, 3102–3113 (2020).
47. Heegaard NHH, Schetter AJ, Welsh JA, Yoneda M, Bowman ED, Harris CC. Circulating micro-RNA expression profiles in early stage nonsmall cell lung cancer. *International Journal of Cancer* **130**, 1378–1386 (2012).
48. Zhang L, *et al.* A High-Performing Plasma Metabolite Panel for Early-Stage Lung Cancer Detection. *Cancers* **12**, (2020).
49. Okamura K, Takayama K, Izumi M, Harada T, Furuyama K, Nakanishi Y. Diagnostic value of CEA and CYFRA 21 – 1 tumor markers in primary lung cancer. *Lung Cancer* **80**, 45–49 (2013).
50. Ørom UA, Nielsen FC, Lund AH. MicroRNA-10a Binds the 5'UTR of Ribosomal Protein mRNAs and Enhances Their Translation. *Molecular Cell* **30**, 460–471 (2008).
51. Louis E, *et al.* Detection of Lung Cancer through Metabolic Changes Measured in Blood Plasma. *Journal of Thoracic Oncology* **11**, 516–523 (2016).
52. Yu T, Yan MX, Lin HC, Liu Y, Chu DD, Tu H, Gu AQ, Yao M. MiRNA-10a is upregulated in NSCLC and may promote cancer by targeting PTEN. *Oncotarget* **6**, 30239–30250 (2015).
53. Liao X, *et al.* LAMP3 regulates hepatic lipid metabolism through activating PI3K/Akt pathway. *Molecular and Cellular Endocrinology* **470**, 160–167 (2018).
54. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).
55. Liu Q, *et al.* Lung regeneration by multipotent stem cells residing at the bronchioalveolar-duct junction. *Nat Genet* **51**, 728–738 (2019).
56. Chow KH, *et al.* S100A4 Is a Biomarker and Regulator of Glioma Stem Cells That Is Critical for Mesenchymal Transition in Glioblastoma. *Cancer Res* **77**, 5360–5373 (2017).
57. Li S, *et al.* Interferon alpha-inducible protein 27 promotes epithelial-mesenchymal transition and induces ovarian tumorigenicity and stemness. *J Surg Res* **193**, 255–264 (2015).
58. Padmanaban V, *et al.* E-cadherin is required for metastasis in multiple models of breast cancer. *Nature* **573**, 439–444 (2019).
59. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer* **12**, 252–264 (2012).

60. Thommen DS, *et al.* Progression of Lung Cancer Is Associated with Increased Dysfunction of T Cells Defined by Coexpression of Multiple Inhibitory Receptors. *Cancer Immunol Res* **3**, 1344–1355 (2015).
61. Puchalska P, Crawford PA. Multi-dimensional Roles of Ketone Bodies in Fuel Metabolism, Signaling, and Therapeutics. *Cell Metab* **25**, 262–284 (2017).
62. Yong WH, *et al.* Ribosomal Proteins RPS11 and RPS20, Two Stress-Response Markers of Glioblastoma Stem Cells, Are Novel Predictors of Poor Prognosis in Glioblastoma Patients. *PLoS One* **10**, e0141334 (2015).
63. Travis WD BE, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, Chirieac LR, Dacic S, Duhig E, Flieder DB, Geisinger K, Hirsch FR, Ishikawa Y, Kerr KM, Noguchi M, Pelosi G, Powell CA, Tsao MS, Wistuba I; WHO Panel. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J Thorac Oncol* **10**, 1243–1260 (2015).
64. Subramanian A, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).
65. Korotkevich GS, Vladimir Sergushichev, Alexey. Fast gene set enrichment analysis. *bioRxiv* **doi**: **https**://**doi**.**org**/**10**.**1101**/**060012**, (2019).
66. Qiu X, *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982 (2017).
67. Iacono G, Massoni-Badosa R, Heyn H. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol* **20**, 110 (2019).
68. Efremova M, Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* **15**, 1484–1506 (2020).
69. Vento-Tormo R, *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).

Tables

Tables 1 - 4 are available as downloads in the Supplementary Files.

Figures

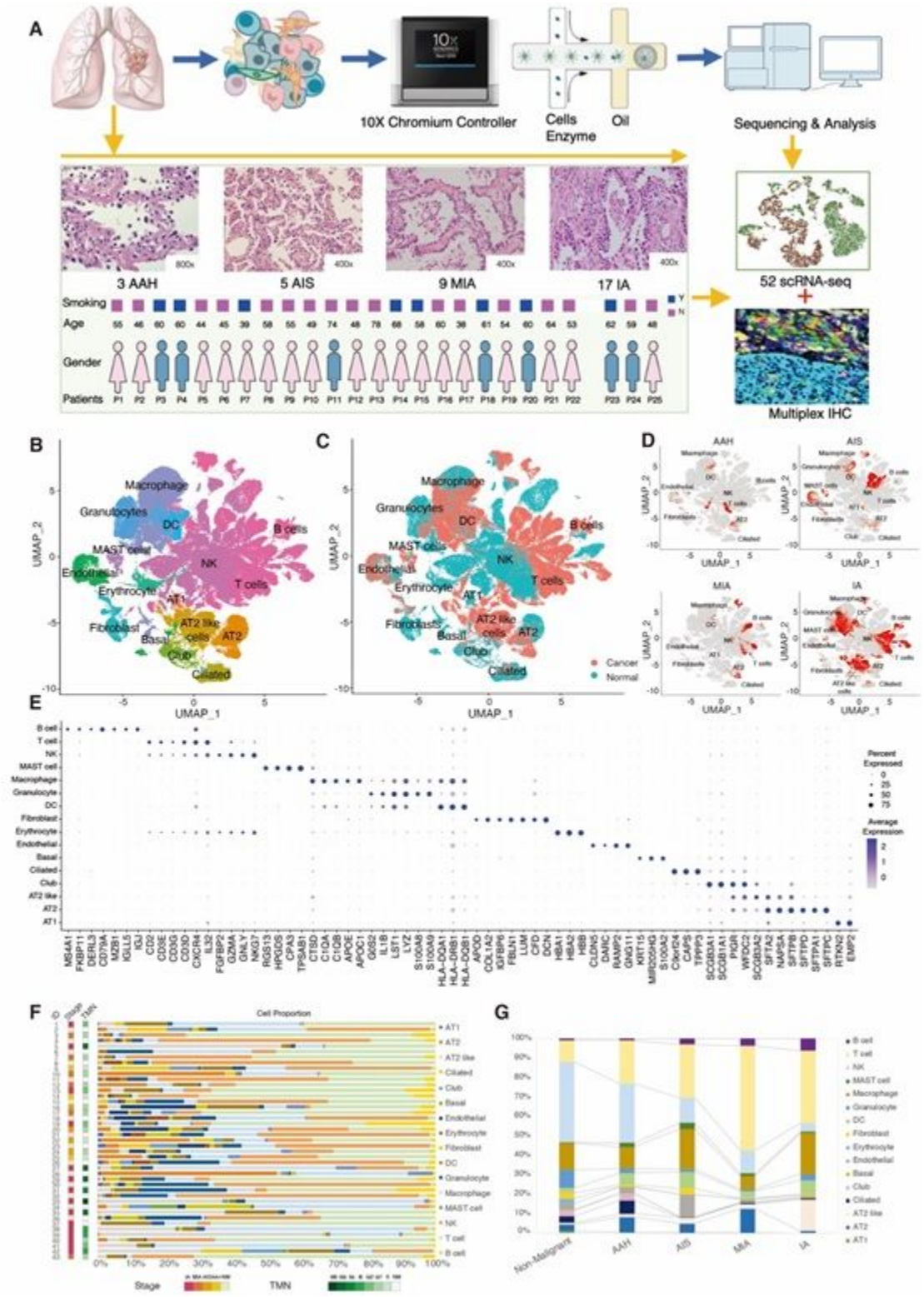

Figure 1

Overview of the single cell transcriptomic profile of LUAD samples. (A). Schematic experimental workflow for the study. A total of 52 freshly resected lung tumor specimens were collected from 25 patients, and non-malignant lung tissues were collected from 18 patients as controls. Samples from patients 1 to 22 were processed using 10X chromium V2 kits, while samples from patients 23 to 25 were used as the validation dataset and processed using 10X Chromium V3 kits. (B). UMAP visualization of 16 major cell

types identified and color-coded by their associated clusters. (C). UMAP visualization of the 16 major cell types identified and color-coded by cancer/normal distribution. (D). UMAP visualization of 16 major cell types identified and color-coded by clinicopathological stages. (E). Dot plot showing the marker genes of each cell type clusters defined in Fig.1B. (F). Proportion of cells originating from different patient samples in the 16 major cell types identified. (G). The percentage of each identified cell type in the different clinicopathological stages.

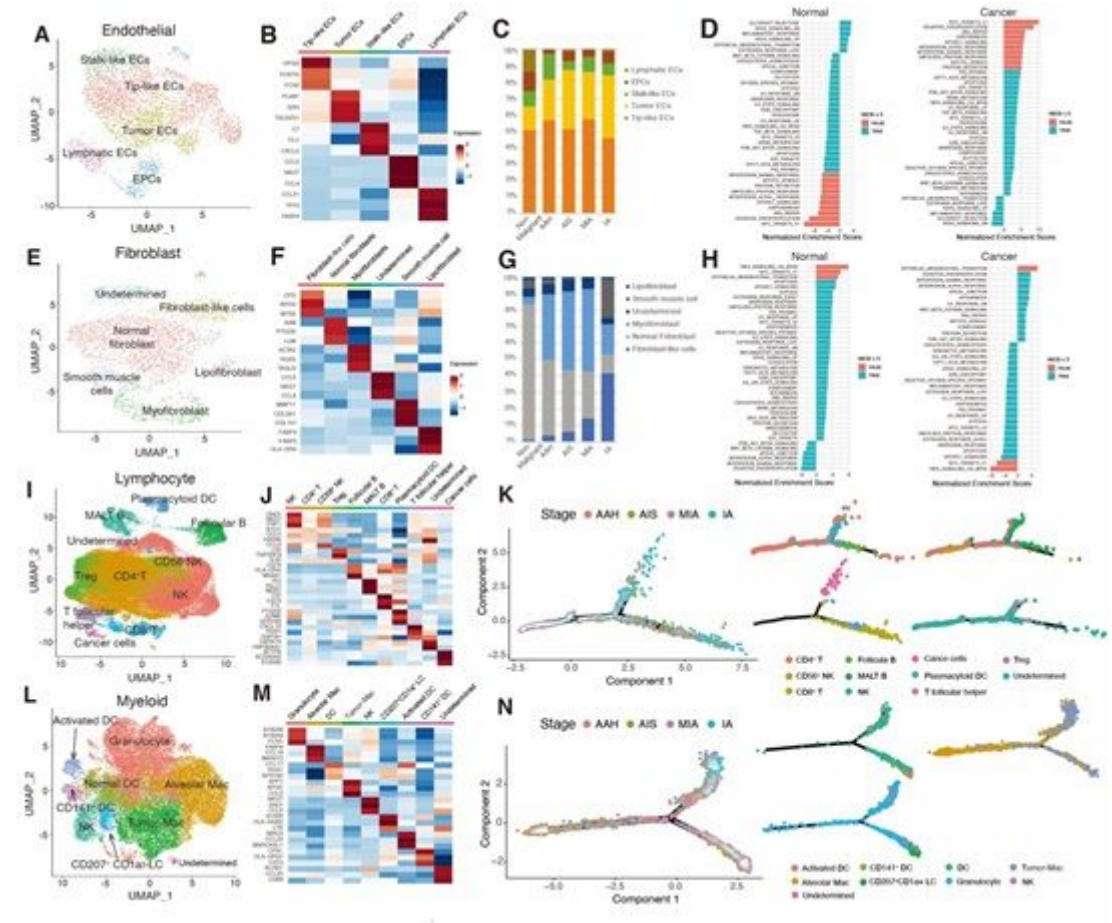

Figure 2

Profiling stromal and immune cell populations in LUAD progression. (A). UMAP visualization of endothelial cell clusters color-coded by identified cell subtypes. ECs: endothelial cells; EPCs: endothelial progenitor cells. (B). Heatmap of selected marker genes in EC cluster subtypes. (C). The percentage of cells within each EC cluster present in the different clinicopathological stages. (D). Top enriched pathway of ECs isolated from normal or tumor tissues as determined by GSEA. (E). UMAP visualization of fibroblast clusters color-coded by identified cell subtypes. (F). Heatmap of selected marker genes in the fibroblast cluster subtypes. (G). The percentage of cells from each fibroblast cluster present in the different clinicopathological stages. (H). Top enriched pathway of fibroblasts isolated from normal or tumor tissues as determined by GSEA. (I). UMAP visualization of lymphocyte clusters color-coded by identified cell subtypes. (J). Heatmap of selected marker genes in the lymphocyte cluster subtypes. (K). Trajectory analysis of lymphocyte clusters. Cells are color coded by clinicopathological stages or identified cell cluster subtypes. (L). UMAP visualization of myeloid cell clusters color-coded by identified

cell subtypes. (M). Heatmap of selected marker genes in the myeloid cell cluster subtypes. (N). Trajectory analysis of myeloid cell clusters. Cells are color-coded by clinicopathological stages or identified cell cluster subtypes.

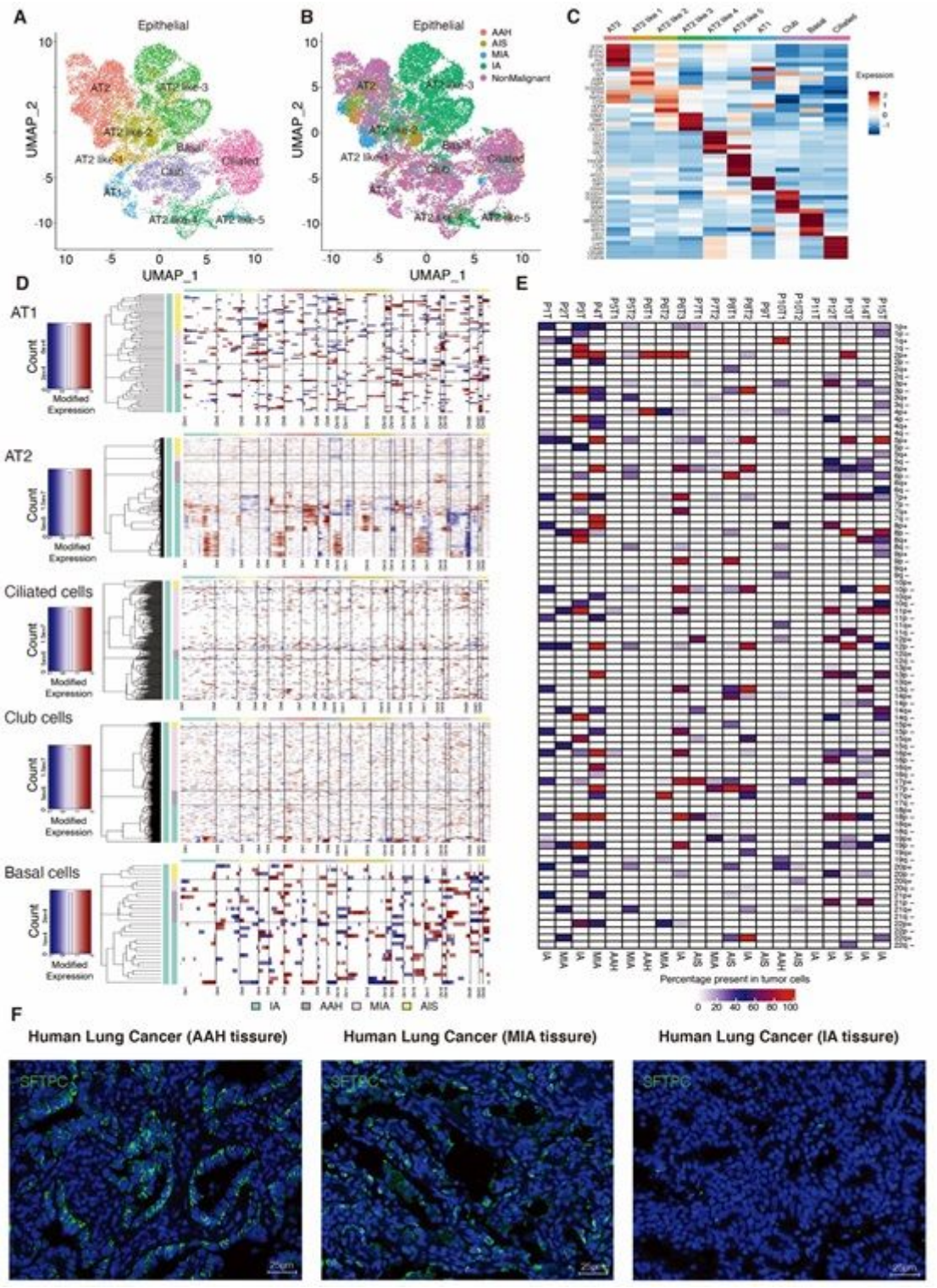

Figure 3

Characterization of epithelial cell lineages across different stages. (A). UMAP visualization of epithelial cell sub-clustering analysis, which is color-coded by the identified cell subtypes. (B). UMAP visualization

of epithelial cell sub-clustering by clinicopathological stage. (C). Heatmap of selected marker genes in each cell cluster subtype. (D). Representative CNV heatmaps with hierarchical clustering of AT1, AT2, ciliated, club, and basal cells. (E). Summary plot of the inferred CNV profiles from each of the 15 patients; CNVs were annotated by the chromosome arm in which the CNV events were calculated. Chromosomal amplification (red) and deletion (blue) are extrapolated in each chromosomal position (columns) across the single cells (rows) using the PDEC protocol. The color bar represents the assigned cell type signature for each cell. (F). Protein fluorescent immunostaining for SFTPC in human tumor samples from the four representative tumor stages (i.e., AAH to IA). Arrows indicate representative SFTPC protein expression in different slices. Scale bars: 50μm; nuclei (DAPI) are stained blue.

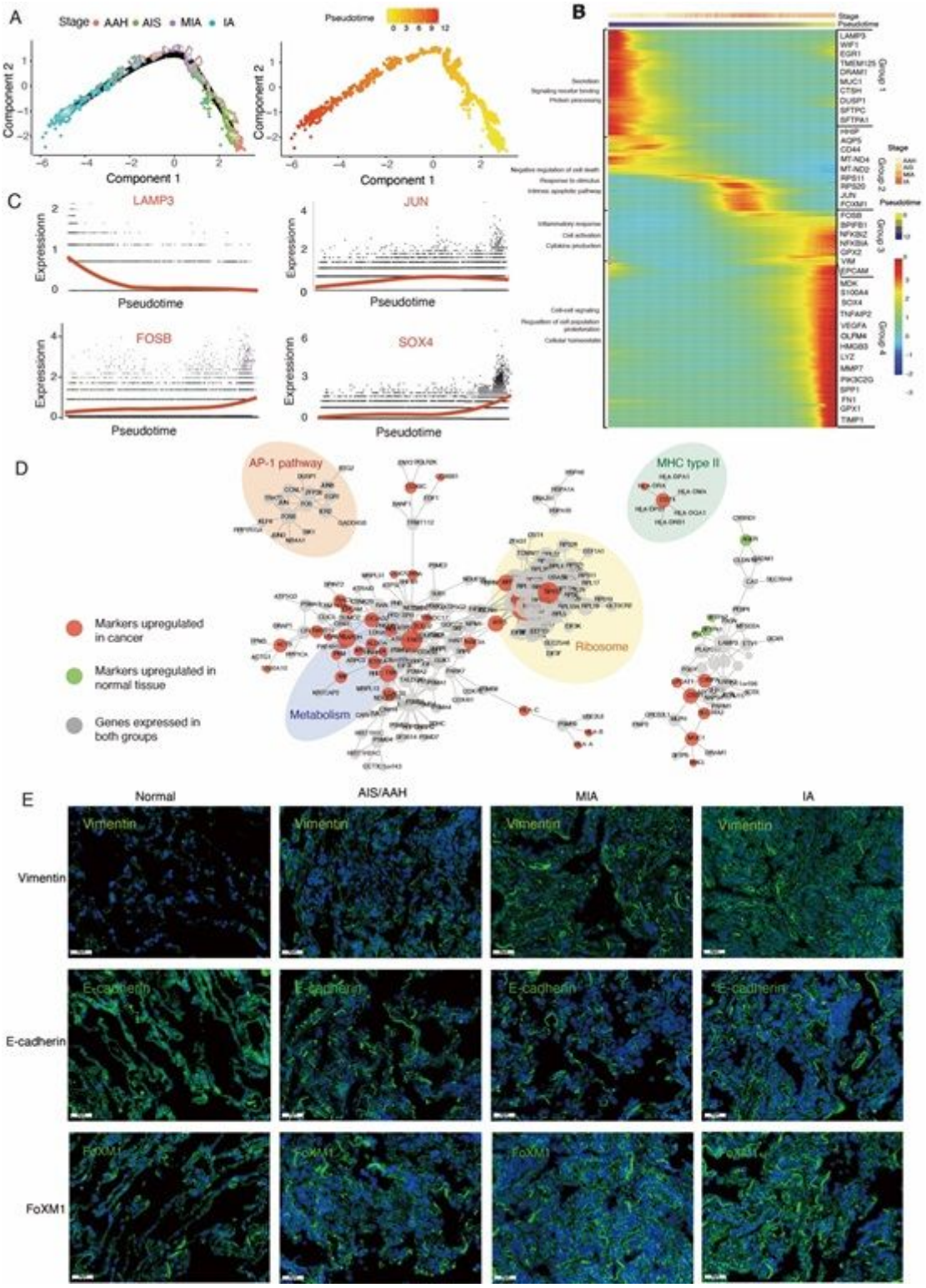

Figure 4

Transcriptional trajectory analysis of AT2 cells. (A). Trajectory analysis of AT2 and AT2-like cell clusters. Cells are color-coded by clinicopathological stages or pseudotime. (B). Heatmap showing 283 differentially expressed genes arranged in pseudotemporal patterns. GO terms from the enrichment analysis reveal the biological function of cells in the 4 groups indicated. (C). Representative gene expression levels of different marker genes, the size of each dot represents relative expression levels. (D).

Gene-gene interaction networks between marker genes in epithelial cell clusters. (E). Fluorescent in situ hybridization staining for vimentin, E-cadherin, and Foxm1 in normal tissues and tumor tissues from different stages. Scale bars: 50μm; nuclei (DAPI) are stained blue.

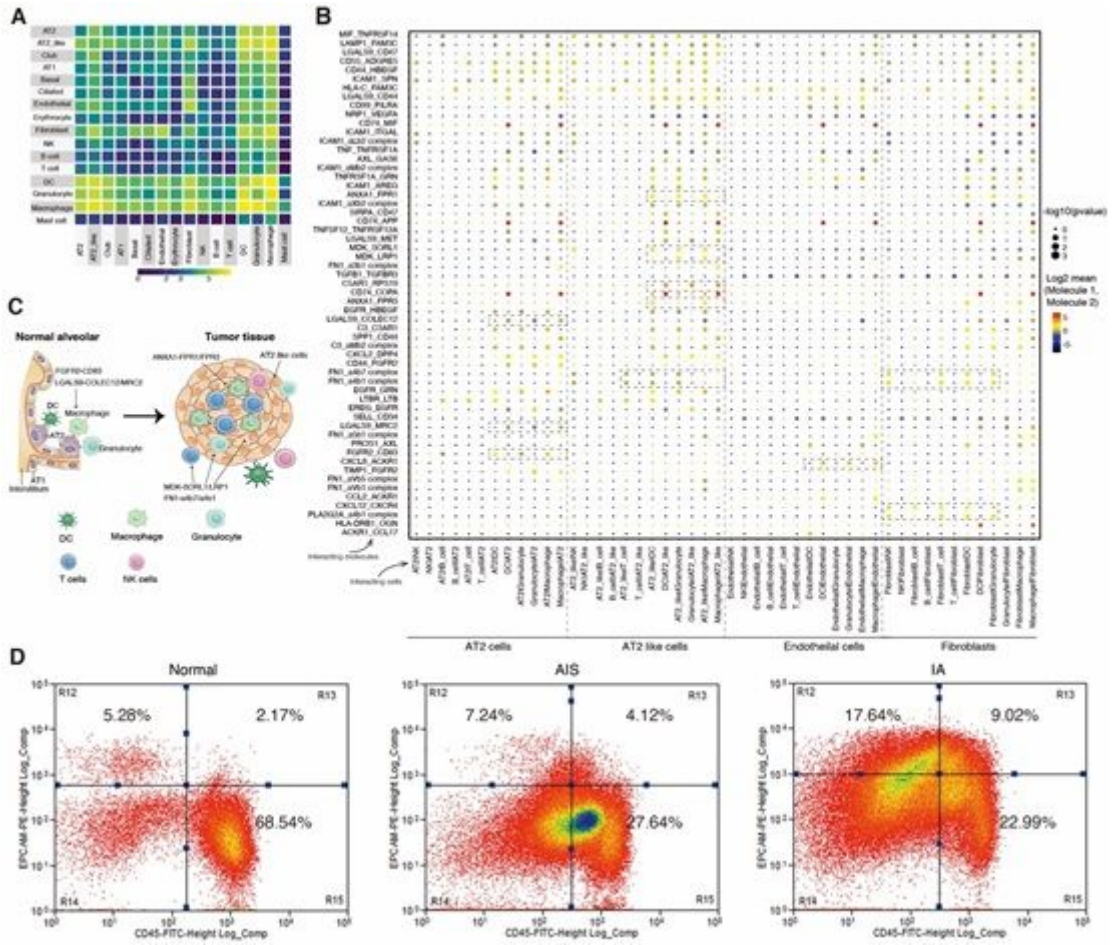

Figure 5

Cell-cell crosstalk visualizing potential specific interactions in LUAD. (A). Heat map depicting the significant interactions among the 16 major cell types identified in Fig.1B. (B). Overview of the selected ligand-receptor interactions; P values are indicated by circle size, scale on right (permutation test, see Methods). The means of the average expression level of interacting molecule 1 in cluster 1 and interacting molecule 2 in cluster 2 are indicated by color. Assays were carried out at the RNA level, but extrapolated to protein interactions. Selected cells include AT2 cells, AT2-like cells, ECs and fibroblasts. (C). Diagram of the main receptors and ligands expressed on AT2 cells and AT2-like cells. (D). Representative flow cytometry plots showing the percentage of Epcam-CD45+ cells in tumor tissues from different stages.

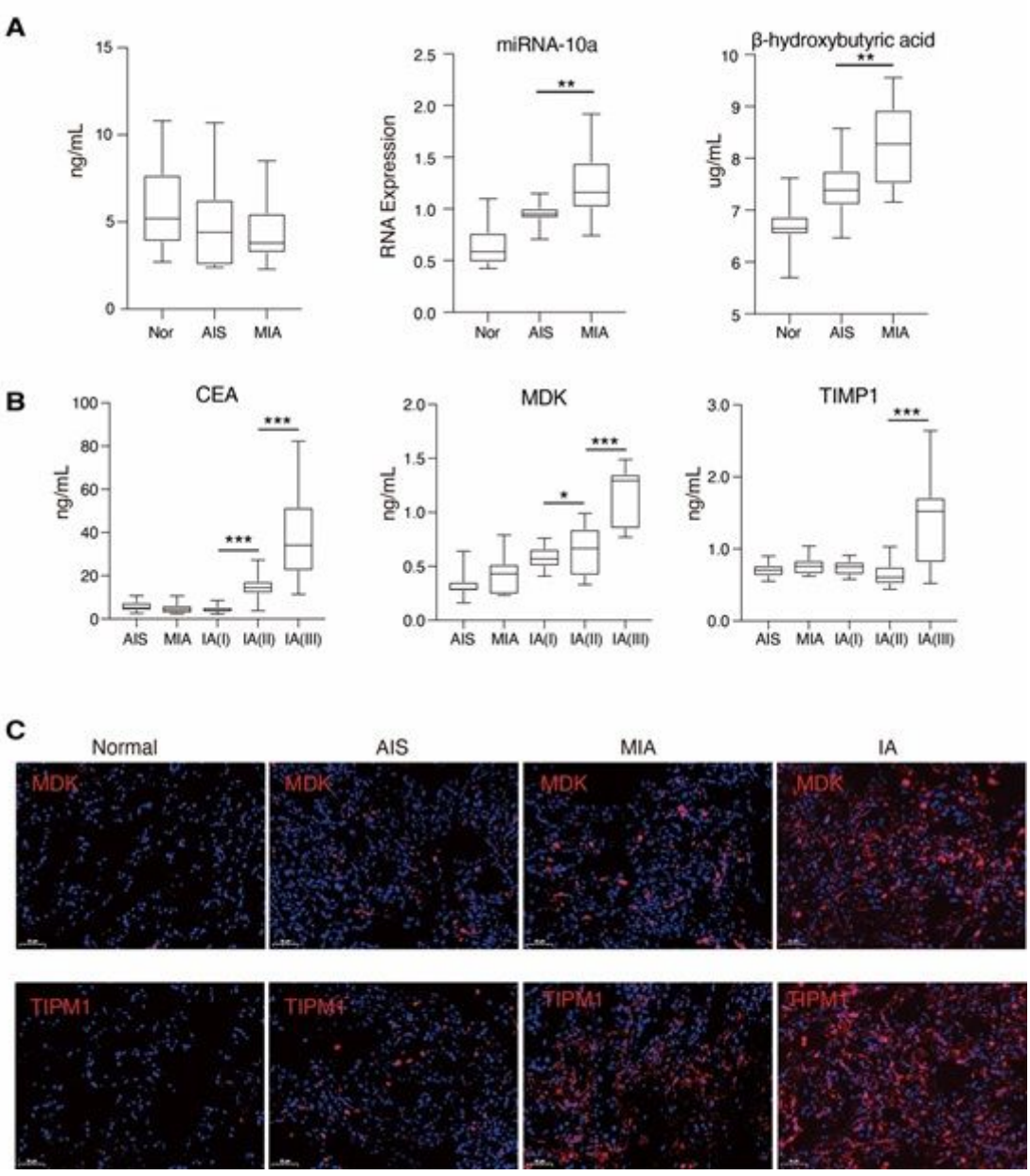

Figure 6

Quantitative detection of key biomarkers in plasma samples from patients with different stages of the LUAD by ELISA or RT-PCR. (A). Detection of mRNA (CEA), miRNA (miRNA-10a) and metabolomics (β-hydroxybutyric acid) in plasma from patients in early-stage LUAD. (B). Detection of protein (CEA, MDK and TIMP1) in plasma from all LUAD patients. n=3; *P < 0.05, **P < 0.01, ***P < 0.001. (C). Protein fluorescent immunostaining for MDK (top), and TIMP1 (bottom) in normal tissues and tumor tissues from different stages. Scale bars: 50μm; nuclei (DAPI) are stained blue.

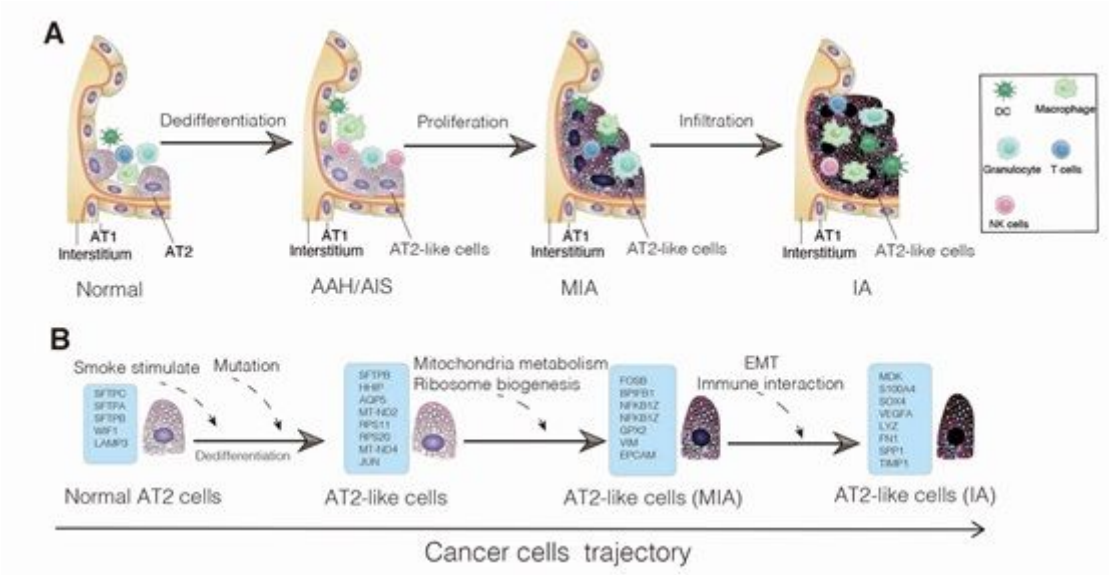

Figure 7

Schematic overview of the proposed cellular (A) and molecular (B) mechanisms involved in the cancer progression of LUAD.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx
- Table1.xlsx
- Table2.xlsx
- Table3.pdf
- Table4.pdf