nature medicine

RESOURCE

https://doi.org/10.1038/s41591-018-0096-5

Phenotype molding of stromal cells in the lung tumor microenvironment

Diether Lambrechts[1,2*], Els Wauters[3,4], Bram Boeckx[1,2], Sara Aibar[5,6], David Nittner[7,8], Oliver Burton[6,9], Ayse Bassez[1,2], Herbert Decaluwé[10,11], Andreas Pircher[1,12], Kathleen Van den Eynde[13], Birgit Weynand[13], Erik Verbeken[13], Paul De Leyn[11], Adrian Liston[6,9], Johan Vansteenkiste[3,4], Peter Carmeliet[1,12,14], Stein Aerts[5,6] and Bernard Thienpont[1,15*]

Cancer cells are embedded in the tumor microenvironment (TME), a complex ecosystem of stromal cells. Here, we present a 52,698-cell catalog of the TME transcriptome in human lung tumors at single-cell resolution, validated in independent samples where 40,250 additional cells were sequenced. By comparing with matching non-malignant lung samples, we reveal a highly complex TME that profoundly molds stromal cells. We identify 52 stromal cell subtypes, including novel subpopulations in cell types hitherto considered to be homogeneous, as well as transcription factors underlying their heterogeneity. For instance, we discover fibroblasts expressing different collagen sets, endothelial cells downregulating immune cell homing and genes coregulated with established immune checkpoint transcripts and correlating with T-cell activity. By assessing marker genes for these cell subtypes in bulk RNA-sequencing data from 1,572 patients, we illustrate how these correlate with survival, while immunohistochemistry for selected markers validates them as separate cellular entities in an independent series of lung tumors. Hence, in providing a comprehensive catalog of stromal cells types and by characterizing their phenotype and co-optive behavior, this resource provides deeper insights into lung cancer biology that will be helpful in advancing lung cancer diagnosis and therapy.

Tumors are characterized by extensive heterogeneity, but so far efforts in understanding this heterogeneity were largely limited to cancer cells[1]. These revealed a remarkably complex and diverse portrait of cancer cells, with evidence for genetic diversification and clonal selection. However, the stromal cells associated with tumors, and the complex cellular ecosystem they build to form the TME, may themselves be as complex and heterogeneous as the cancer cell compartment[2,3]. Particularly, an increasing number of studies suggest that stromal cells, such as macrophages, T cells and fibroblasts, are highly heterogeneous[4–7]. The extent of this heterogeneity, how it is shaped by other cells in the tumor and vice versa also directly affects them, remains however poorly characterized, in part because of a historical lack of methods to study these cells in isolation.

Notwithstanding these open questions, the TME is increasingly recognized as a cancer therapy target. Non-small-cell lung cancer (NSCLC) above all seems to benefit from such novel treatments. For instance, antibodies targeting the programmed cell death-1 receptor (PD-1) or ligand (PD-L1) activate antitumoral responses of cytotoxic T cells. In advanced NSCLC patients, these treatments demonstrated response rates up to 45%, with some responses being remarkably durable[8,9]. Likewise, the triple angiokinase inhibitor nintedanib, when added to docetaxel, significantly extends median overall survival in previously-treated NSCLC patients[10]. Intriguingly, despite the paramount therapeutic importance, the in situ phenotype of stromal cells targeted remains elusive.

The advent of single-cell RNA-sequencing (scRNA-seq) enables specific profiling of cell populations at the single-cell level. While conventional 'bulk' RNA-sequencing (RNA-seq) methods process millions of cells, averaging out underlying differences, scRNA-seq can reveal changes that render each individual cell type unique. Moreover, advances in microfluidics enable simultaneous profiling of thousands of cells from a biopsy sample[11]. This allows unbiased assessment of many heterogeneous stromal and cancer cells at the single-cell level, hence revealing complexities of the molecular components and differences with counterparts residing in non-malignant tissue. Previous scRNA-seq studies on glioblastoma[1], melanoma[12] and oligodendroglioma[12] focused largely on cancer cells, analyzing few stromal cells from tumors, and not from matching non-malignant tissue. By analyzing cells from tumors and matching non-malignant tissue at a much higher scale, we uncover stromal cell heterogeneity and adaptation to the tumor.

Results

scRNA-seq and cell typing of non-malignant lungs and lung tumors. Five patients with untreated, non-metastatic NSCLC of the squamous cell (lung squamous carcinoma (LUSC)) or adenocarcinoma subtype (lung adenocarcinoma (LUAD)) underwent

[1]VIB Center for Cancer Biology, Leuven, Belgium. [2]Laboratory for Translational Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium. [3]Respiratory Oncology Unit (Pneumology) and Leuven Lung Cancer Group, University Hospitals KU Leuven, Leuven, Belgium. [4]Laboratory of Pneumology, Department of Chronic Diseases, Metabolism and Ageing, KU Leuven, Leuven, Belgium. [5]Laboratory for Computational Biology, Department of Human Genetics, KU Leuven, Leuven, Belgium. [6]VIB-KU Leuven Center for Brain & Disease Research, Leuven, Belgium. [7]Histopathology Expertise Center, VIB Leuven Center for Cancer Biology, VIB, Leuven, Belgium. [8]Department of Oncology, KU Leuven, Leuven, Belgium. [9]Laboratory of Genetics of Autoimmunity, Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium. [10]Department of Thoracic Surgery, University Hospitals KU Leuven, Leuven, Belgium. [11]Department of Chronic Diseases, Metabolism and Ageing, KU Leuven, Leuven, Belgium. [12]Laboratory of Angiogenesis and Vascular Metabolism, Department of Oncology, KU Leuven, Leuven, Belgium. [13]Translational Cell & Tissue Research, Department of Imaging & Pathology, KU Leuven, Leuven, Belgium. [14]State Key Laboratory of Ophthalmology, Zhongsan Ophthalmic Center, SunYat-Sen University, Guangzhou, China. [15]Laboratory for Functional Epigenetics, Department of Human Genetics, KU Leuven, Leuven, Belgium. *e-mail: diether.lambrechts@kuleuven.vib.be; bernard.thienpont@kuleuven.be

lung lobe resection with curative intent. All patients were former smokers and some had mild chronic obstructive pulmonary disease (COPD) (Fig. 1a, Supplementary Table 1). Following resection, one non-malignant lung tissue sample from a distal region within the same lobe and three tumor tissue samples were obtained, rapidly digested to a single-cell suspension and analyzed using scRNA-seq involving a single-tube protocol with unique transcript counting through barcoding with unique molecular identifiers (UMIs) (see Methods). After quality filtering (see Methods), ~0.2 billion unique transcripts were obtained from 52,698 cells in which over 100 genes could be detected as expressed. Of these, 39,323 cells (75%) originated from lung tumors and 13,375 from non-malignant lungs (Fig. 1a,b, Supplementary Table 2). Following gene expression normalization for read depth and mitochondrial read count, we applied principle component analysis on genes variably expressed across all 52,698 cells ($n=2,192$ genes). Subsequently, we classified cells into groups of cell types using graph-based clustering on the informative principle components ($n=8$). This identified cell clusters that, through marker genes, could be readily assigned to known cell lineages: in addition to cancer cells, we identified immune cells (myeloid, T and B cells), fibroblasts, endothelial cells, alveolar cells and epithelial cells (Fig. 1c; Supplementary Fig. 1). These differed considerably in transcriptional activity, as we detected on average 1,678 transcripts (764 genes) per T cell and 6,746 transcripts (1,828 genes) per cancer cell (Fig. 1d; Supplementary Table 3).

To corroborate these profiles, we in parallel also performed bulk RNA-seq of a tumor and non-malignant sample. Notably, no pronounced effects of cell dissociation on gene expression[13] were noted and bulk transcript counts correlated well with scRNA-seq data ($r=0.71$; Supplementary Fig. 2a,b). Ontology analysis of differentially expressed gene sets revealed that they were enriched (immune-related processes) or depleted (epithelium, extracellular matrix) in the scRNA-seq data (Supplementary Fig. 2c). Cell cluster marker gene expression was also dissimilar (Supplementary Fig. 2b). These differences may reflect well-known disparities in dissociation efficiency of different cell types following tissue disaggregation[14], with fibroblasts and endothelial cells being more embedded in extracellular matrix and basement membrane than immune cells, and hence more difficult to dissociate. To estimate relative contributions of each cell cluster to bulk RNA expression, we applied quadratic programming, which confirmed enrichment of alveolar cells and lymphocytes (B and T cells) and a paucity of fibroblasts and endothelial cells in the tumor and non-malignant sample (Supplementary Fig. 2d,e). Hence, while scRNA-seq faithfully reproduces bulk expression profiles, differences in single-cell dissociation efficiency influence recovery of individual cell types.

To identify subclusters within each of these eight major cell types, we performed principle component analysis within each cell type (see Methods). For this analysis, we validated clustering robustness by varying parameter settings, including resolution, *k*-means and number of informative principle components, and selected the most robust method amongst five clustering methods (Seurat[11]; Supplementary Fig. 3). Notably, as very few cells were positive for cell proliferation markers, we opted not to correct for cell cycle (Supplementary Fig. 4). Overall, this analysis revealed the presence of a complex cellular ecosystem, containing 52 different stromal cell subclusters and 12 cancer cell subclusters (Fig. 1d). Importantly, when comparing between patients, cancer cell subclusters were highly patient-specific (Fig. 1d), consistent with somatic mutations being tumor- or patient-specific. In contrast, stromal cell subclusters mostly consisted of cells from three or more patients. A validation cohort of a further 40,250 single cells from 3 additional NSCLC patients (Supplementary Table 1) revealed that stromal cells could be assigned to 45 of 52 subclusters, representing 86% of stromal cells in the original set of 5 patients (Supplementary Fig. 5). Moreover, we recovered >10 cells from both LUSC and LUAD tumors for 46 of 52 subclusters. Together, this argues against strong interindividual variation of stromal cells and suggests that these 52 cell subtypes cover most of the cellular heterogeneity in the lung TME. Strikingly, when comparing tumors and matching non-malignant lungs, many stromal cell subclusters were enriched for either tumor-derived or lung tissue-derived cells (Fig. 1d; Supplementary Fig. 6). Such enrichment was replicated in the additional three NSCLC patients ($r=0.77$, $P<10^{-5}$; Supplementary Fig. 5). We therefore explored these changes in greater detail for the main stromal cell types.

Tumor endothelial cells downregulate immune attraction pathways. We detected 1,592 endothelial cells. As expected given the hypervascular nature of lungs, endothelial cells were less abundant in the tumor[15]. Reclustering these 1,592 endothelial cells revealed 6 clusters (Fig. 2a). We next attempted to identify marker genes for each of these clusters and to assign them to known endothelial cell types (Supplementary Table 3). This revealed one set of 85 lymphatic endothelial cells found in tumor and non-malignant samples (cluster 6; marker genes *PDPN* and *PROX1*), and 5 sets of blood endothelial cells (*FLT1*+; Fig. 2b): two were mostly tumor-derived (clusters 3 and 4; *IGFBP3*+and *SPRY1*+) and two others were mostly non-malignant lung-derived (clusters 1 and 5; *MT2A*+ and *EDNRB*+) (Fig. 2a,b; Supplementary Fig. 6). The remaining cluster contained lower quality endothelial cells and was disregarded for further analyses, although biological functions cannot formally be excluded (cluster 2; no marker genes). A similar enrichment was observed in 40,250 cells from 3 additional patients (Supplementary Fig. 5). Likewise, when assessing expression of marker genes in bulk RNA-seq from 108 non-malignant lungs, 501 LUSC or 513 LUAD tumors cataloged in The Cancer Genome Atlas (TCGA), marker genes for normal and tumor endothelial cells, were enriched in non-malignant lungs and lung tumor, respectively (Fig. 2c). Immunofluorescence analysis of independent lung tumors and non-malignant samples for ACKR1 and EDNRB, markers of cluster 3 and 5, respectively, confirmed presence of these cells as separate cellular entities, respectively enriched in tumor and non-malignant tissue (Supplementary Fig. 7).

Analysis of hallmark pathway gene signatures[16] highlighted that, while the two tumor endothelial cell clusters showed some differences (Supplementary Fig. 8a), most changes were between non-malignant lung and tumor-derived endothelial cells. A direct comparison of tumor versus normal endothelial cells revealed Myc targets as the top enriched signature in tumor endothelial cells (Fig. 2d). Remarkably, total read counts in tumor endothelial cell clusters were two- to fourfold higher than in normal endothelial cell clusters. This was not due to PCR bias artefacts or altered expression of RNA-degradation enzymes (Supplementary Fig. 9), suggesting that tumor endothelial cells have a higher RNA content due to increased rates of transcription (Fig. 2e). Because Myc can almost universally upregulate transcription[17], this suggests Myc to underlie this effect, identifying a potential vulnerability of tumor endothelial cells to Myc inhibition. Indeed, earlier studies indicated that c-Myc is essential for tumor angiogenesis[18]. Metabolic pathway analysis further supported this increased requirement for transcription, as the most significantly increased pathway was involved in nucleotide metabolism (purine and pyrimidine biosynthesis; Supplementary Fig. 8b–d). Other metabolic pathways affected include those involved in oxidative phosphorylation and glycolysis, which are instrumental for angiogenesis[19–21]. Therapeutically inhibiting glycolysis in endothelial cells moreover normalizes tumor angiogenesis in mice[19].

Surprisingly, the most significantly downregulated pathway was involved in inflammatory responses. A more detailed analysis revealed downregulation of genes involved in immune activation and immune cell homing (Fig. 2f). Importantly, the endothelium represents the primary interface between circulating immune cells and the tumor, and plays important roles in relaying signals and

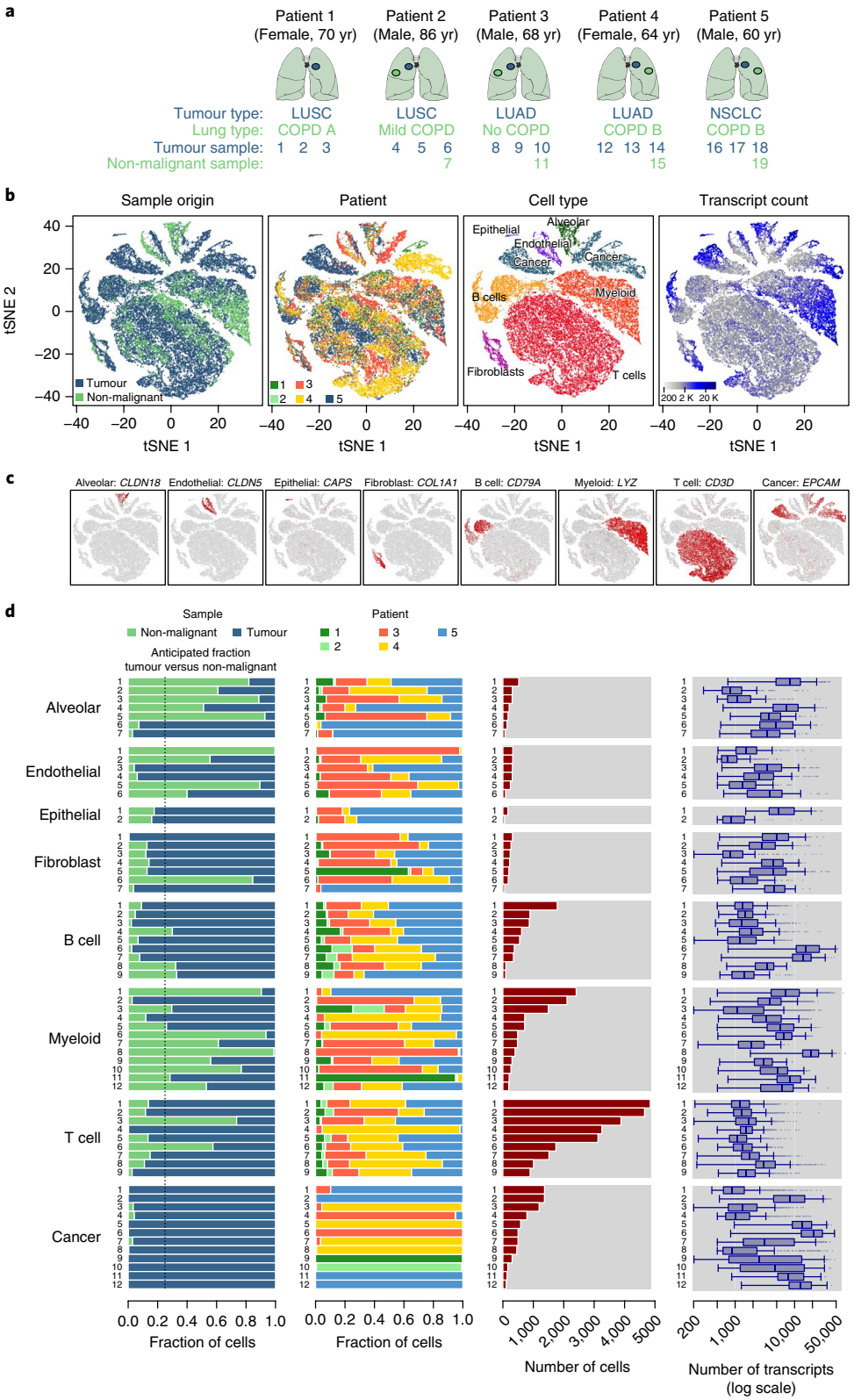


Fig. 1 | Overview of the 52,698 single cells from lung tumors and distal non-malignant lung samples. a, Summary of the sample origins. COPD classification according to Global Initiative for Chronic Obstructive Lung Disease staging. **b**, tSNE of the 52,698 cells profiled here, with each cell color-coded for (left to right): its sample type of origin (tumor or non-malignant lung), the corresponding patient, the associated cell type and the number of transcripts (UMIs) detected in that cell (log scale as defined in the inset). K, thousand. **c**, Expression of marker genes for the cell types defined above each panel. Three additional marker genes for each cell type are shown in Supplementary Fig. 1. **d**, For each of the 52 stromal cell subclusters and the 12 cancer cell subclusters (left to right): the fraction of cells originating from the 4 non-malignant and 15 tumor samples, the fraction of cells originating from each of the 5 patients, the number of cells and box plots of the number of transcripts (with plot center, box and whiskers corresponding to median, IQR and 1.5 × IQR, respectively; *n* per boxplot is shown in the 'number of cells' panel, and specified in Supplementary Table 3).

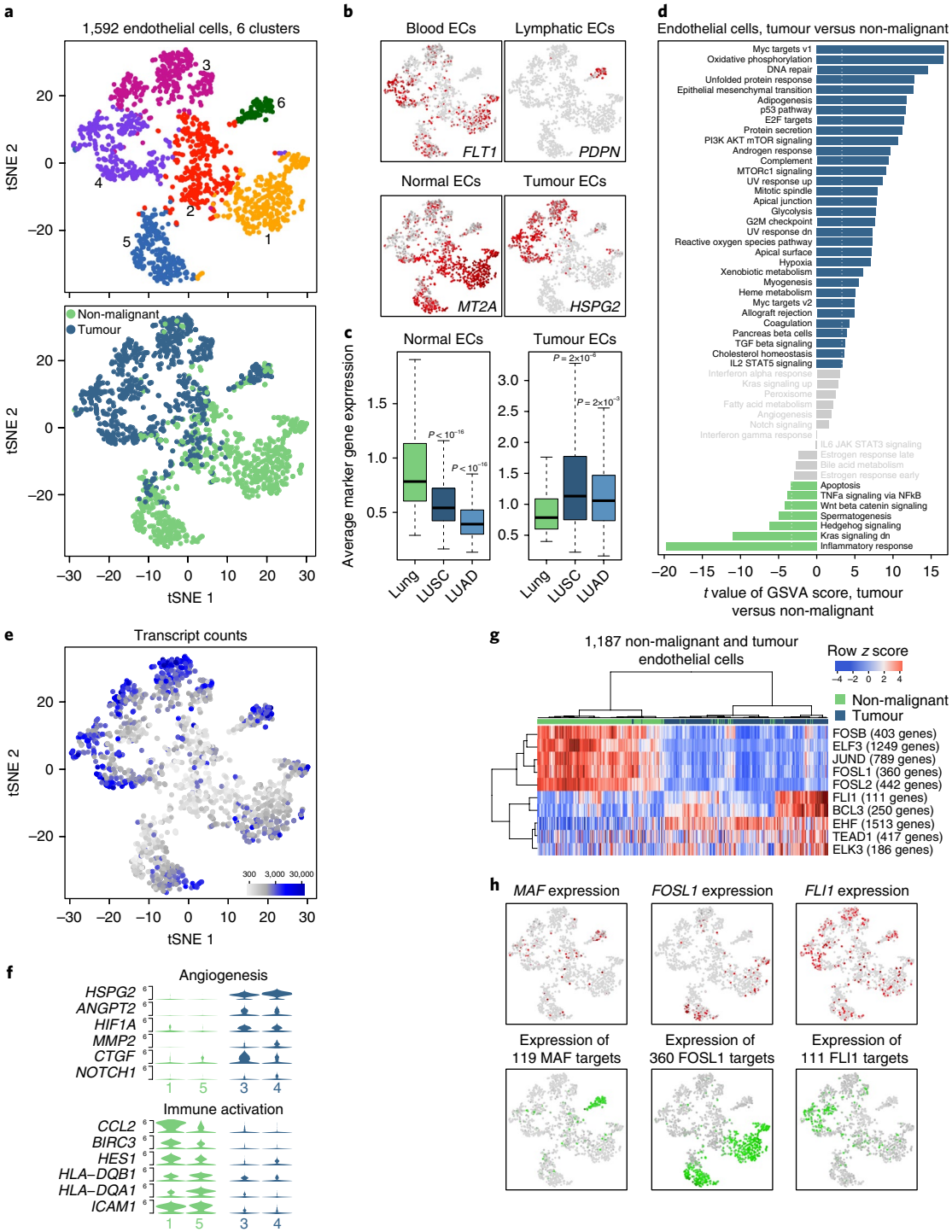


Fig. 2 | Endothelial cell clusters. a, tSNE plot of 1,592 endothelial cells, color-coded by their associated cluster (top) or the sample type of origin (bottom). Note that cluster 2 consists mainly of lower quality cells, and is discarded from further analyses. **b**, tSNE plot color-coded for expression (gray to red) of marker genes for blood, lymphatic, tumor and normal endothelial cells. EC, endothelial cell. **c**, Average expression of 29 marker genes for normal endothelial cells or 7 marker genes for tumor endothelial cells (Supplementary Table 3) in TCGA samples from lung ($n=108$), LUSC ($n=501$) or LUAD ($n=513$). Expression of each gene is normalized to its average expression in non-malignant lung samples. Box plot center, box and whiskers correspond to median, IQR and 1.5×IQR, respectively. Data were analyzed using one-way ANOVA with Tukey's multiple-comparisons test. **d**, Differences in pathway activities scored per cell by GSVA between tumor and normal endothelial cell ($n=618$ and 569 cells from 5 patients, respectively). Shown are *t* values from a linear model, corrected for patient of origin. dn, down; UV, ultraviolet; v1, version 1; v2, version 2. **e**, tSNE plot of endothelial cells, color-coded according to the number of transcripts detected in each cell. **f**, Violin plots showing the smoothened expression distribution of selected genes involved in angiogenesis and immune activation, stratified per normal or tumor endothelial cell cluster (green and blue, $n=323$, 246, 311 and 307 endothelial cells for clusters 1, 5, 3 and 4, respectively). **g**, Heatmap of the area under the curve (AUC) scores of expression regulation by transcription factors, as estimated using SCENIC, for each of the 1,187 endothelial cells from clusters 1, 3, 4 and 5. Shown are the five transcription factors having the highest difference in expression regulation estimates between tumor and normal endothelial cells. **h**, tSNE plots of endothelial cells, color-coded for (top) the expression of (left to right) MAF, FOSL1 and TEAD1, and for (bottom) the AUC of the estimated regulon activity of these transcription factors, corresponding to the degree of expression regulation of their target genes.

presenting epitopes from the tissues it vascularizes to the immune system[22]. Gene classes downregulated included those involved in antigen presentation (major histocompatibility complex class I and II), chemotaxis (*CCL2*, *CCL18*, *IL6*) and immune cell homing (*ICAM1*) (Fig. 2f). Together, this indicates that tumor endothelial cells are remodeled to downregulate their antigen presentation and immune cell homing activities, thus contributing to tumor immunotolerance. These data extend recent findings demonstrating synergistic effects of tumor vessel normalization and checkpoint immunotherapy[23].

Finally, to assess which transcription factors underlie differences in expression between tumor and normal endothelial cells, we applied Single-Cell Regulatory Network Inference And Clustering (SCENIC)[24]. SCENIC scans differentially expressed genes for overrepresented transcription factor binding sites, and analyses coexpression of transcription factors and their putative target genes. This identified MLX and MAF as candidate transcription factors underlying gene expression differences in lymphatic endothelial cells, whereas downregulation of Fos/Jun and ELF3, and upregulation of FLI1 and TEAD1, seemed responsible for tumor-specific endothelial cell phenotypes (Fig. 2g,h). Interestingly, Fos/Jun are linked to *ICAM1* expression in endothelial cells[25], and also other genes involved in immune activation show putative Fos/Jun binding sites, indicating Fos/Jun loss to underlie the reduced immune stimulatory phenotype of tumor endothelial cells.

Lung tumors harbor five distinct types of fibroblasts. Fibroblasts have long been suggested to represent a heterogeneous population but the extent of heterogeneity has hitherto remained unexplored, as fibroblast phenotypes are considered highly context-dependent and unstable in culture[26]. In our samples, 1,465 fibroblasts were detected. Subclustering revealed seven distinct subtypes. While fibroblasts were overall only modestly enriched in tumors, cluster 1 was strongly enriched in tumors and cluster 6 was enriched in non-malignant samples. Most clusters were found in three or more patients (Fig. 1d; Fig. 3a,b), and all but one (cluster 7) were present in three additional NSCLC patients (Supplementary Fig. 5). Enrichment of these clusters in tumor and non-malignant lung was confirmed in these three patients and in bulk RNA-seq from TCGA (Fig. 3c). Immunofluorescence analysis of independent lung tumors and non-malignant lungs for COX4I2 and FIGF, markers of clusters 2 and 6, moreover confirmed presence of these clusters as separate cellular entities (Supplementary Fig. 10).

Remarkably, each of these fibroblast types expresses a unique repertoire of collagens and other extracellular matrix molecules, with for example cluster 1 expressing *COL10A1* and cluster 2 expressing *COL4A1* (Fig. 3c). In contrast to tumor-derived fibroblasts, non-malignant fibroblasts (cluster 6) express high elastin levels and low levels of some collagens (collagens type I, III, V and VIII) but not others (for example, collagen type VI) (Fig. 3d). As different collagens have different roles in the extracellular matrix, this suggests functional specialization of fibroblast clusters[27]. To characterize their functions in greater detail, we compared pathway activities and observed significant phenotypic diversity (Fig. 3e). *ACTA2*, a myofibroblast marker[26], showed highest expression in cluster 2. This cluster also displayed high expression of other genes involved in myogenesis (for example, *MEF2C*, *MYH11* or *ITGA7*), the NOTCH pathway and angiogenesis, suggesting these cells are strongly activated. Notably, pericytes coclustered with cluster 2, as a subset expresses *RGS5*, a pericyte marker[26]. Although clusters 5 and 7 were highly similar, with lower myogenesis and high mTOR signature expression, they differed in expression of glycolysis genes, indicating metabolic differences. Also, clusters 1 and 4 were similar, but cluster 1 showed a strong epithelial-mesenchymal transition signal in line with expression of an extensive repertoire of extracellular matrix proteins and TGF-β-associated genes.

When contrasting SCENIC in cells from one fibroblast cluster versus all other fibroblasts, genes regulated by MEF2C and ELK3 were highly upregulated in cluster 2, while genes regulated by FOXO1 and MSC were downregulated (Fig. 3f). With MEF2C being a known myogenic transcription factor[28] and MSC a myogenic inhibitor[29], these analyses identify plausible candidates for the prominent cluster 2 myogenesis phenotype (Fig. 3g). Likewise, in cluster 1, genes regulated by HOXB2 and FOXO1 were highly upregulated. Genes encoding extracellular matrix proteins such as *COL1A1*, *COL3A1* and *COL6A1* have a particularly high number of putative HOXB2 and FOXO1 binding sites near their promoter; these transcription factors are likely to underlie the extracellular matrix phenotype of cluster 1 fibroblasts (Fig. 3g).

B cells are strongly enriched in the tumor. We detected 4,806 B lymphocyte cells and 797 other cells that cluster near B cells (Supplementary Fig. 11). B cells represent the most tumor-enriched stromal cell type (Fig. 1d; Supplementary Fig. 6). Clustering revealed nine clusters. Of these, six were particularly tumor-enriched: follicular B cells expressing high levels of CD20 (*MS4A1*), CXCR4 and HLA-DRs (clusters 1 and 2), plasma B cells expressing immunoglobulin gamma (clusters 3 and 6) and mucosa-associated lymphoid tissue-derived (MALT) B cells expressing immunoglobulins A and M and JCHAIN (clusters 5 and 7) (Fig. 4a–c; Supplementary Table 3). While plasma B cells do not express cell proliferation markers (Supplementary Fig. 4), we cannot exclude that some are plasmablasts. Pathway analyses failed to identify differences between non-malignant lung-derived and tumor-derived plasma or MALT B cells, although the low number of cells isolated from non-malignant lung may affect the power to identify changes. We did identify differences in follicular B cells, revealing tumor-associated decreases in oxidative phosphorylation, cell proliferation and biomass production (that is, pathways associated with Myc, mTOR and protein secretion) (Fig. 4d). In line with this, transcript numbers were 37.9% lower in tumor-associated versus non-malignant lung-associated follicular B cells (Supplementary Fig. 9). Together, these data suggest that follicular B cells become exhausted in the tumor.

Other cells that coclustered with B cells correspond to immune cells outside of the B-cell lineage (Supplementary Fig. 11): mast cells (cluster 4; positive for tryptases, *KIT* and *MS4A2*), plasmacytoid dendritic cells (cluster 8; *LILRB4*+) and erythroblasts (cluster 9; *HBB*+, *ALAS2*+ and *SNCA*+). These were not particularly tumor-enriched (Fig. 4a–c, Supplementary Fig. 6). They moreover failed to show separation between tumor and non-malignant lungs on t-distributed stochastic neighbor embedding (tSNE), indicating that they are not strongly shaped by the tumor (Fig. 4a). Pathway analyses similarly failed to identify strong differences.

Macrophages show rheostatic phenotypes and become M2 polarised in tumors. The 9,756 myeloid cells clustered in 12 separate subsets (Fig. 4e). One cluster corresponds to granulocytes (cluster 7; *S100A12*+) and three clusters to dendritic cells: Langerhans cells (cluster 5; *CD207*+), monocyte-derived dendritic cells (cluster 9; *FCGR3A*+, *CYTIP*+) and cross-presenting dendritic cells (cluster 12; *CLEC9A*+ and *XCR1*+) (Supplementary Fig. 12). These cell types were typically less abundant in tumor than non-malignant tissue, apart from the Langerhans cells which were detected at similar numbers (Supplementary Fig. 6). The eight other clusters consisted of CD163+ and CD68+ macrophages and were also less abundant in tumor (Fig. 4f; Supplementary Fig. 6) and displayed extensive heterogeneity driven by both patient and tissue specificity: five clusters were either tumor- or non-malignant lung-derived for >85% of cells, while four clusters were >85% derived from one patient (Fig. 1d). Enrichment of these clusters in non-malignant lung or tumor was confirmed by a recent study of the lung tumor immune landscape[7], in 40,250 additional cells profiled by scRNA-seq

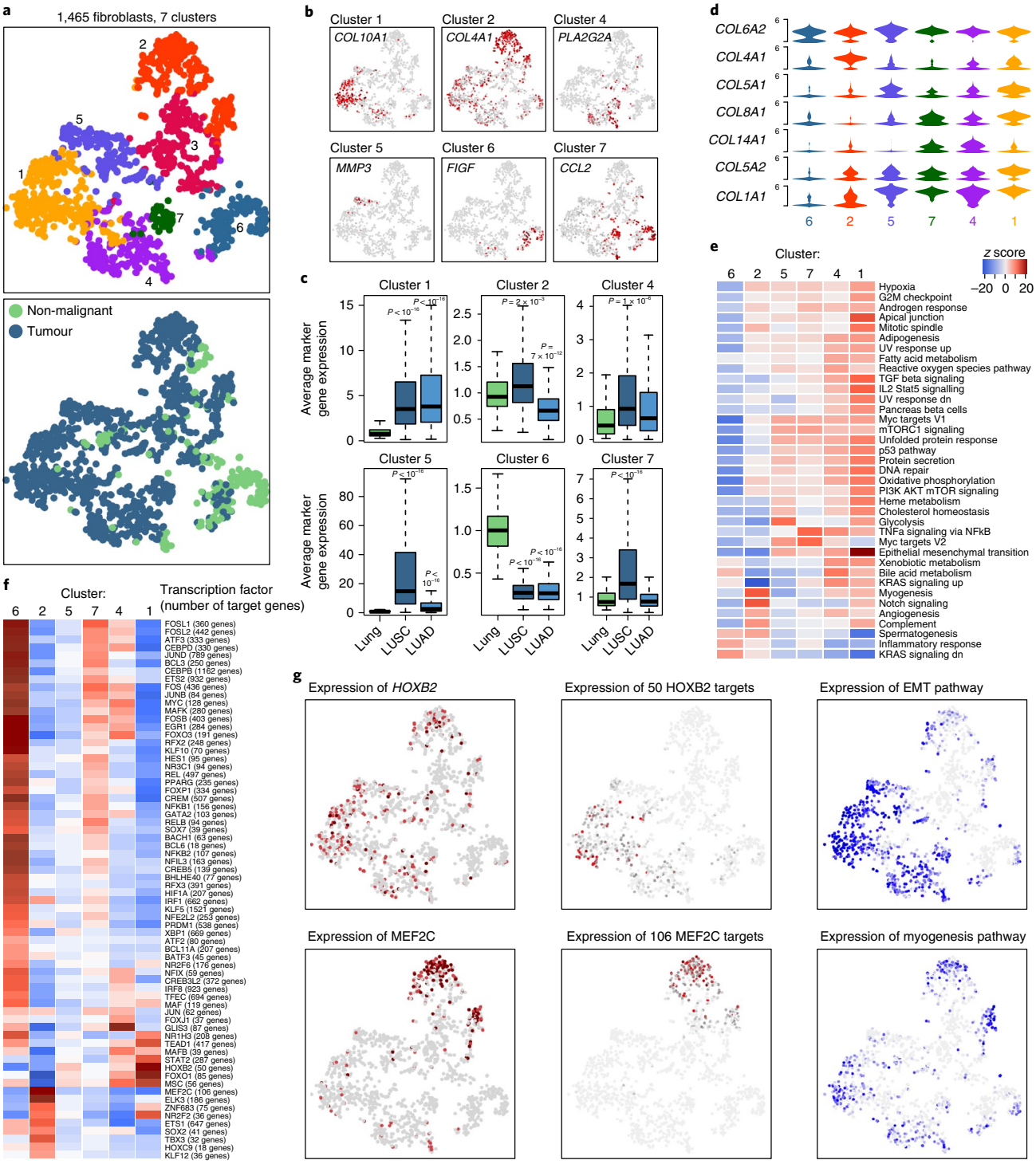


Fig. 3 | Fibroblast clusters in lungs and lung tumors. a, tSNE plot of 1,465 fibroblasts, color-coded by their associated cluster (top) or the sample type of origin (bottom). Note that cluster 3 consists mainly of lower quality cells, and is discarded here from further analyses. **b**, tSNE plot color-coded for expression (gray to red) of marker genes for the clusters as indicated. **c**, Average expression of marker genes for fibroblasts from each cluster (Supplementary Table 3) in TCGA samples from lung ($n=108$), LUSC ($n=501$) or LUAD ($n=513$). Expression of each gene is normalized to its average expression in non-malignant lung samples. Box plot center, box and whiskers correspond to median, IQR and 1.5 × IQR, respectively. Data were analyzed using one-way ANOVA with Tukey's multiple-comparisons test. **d**, Violin plots showing the smoothened expression distribution of selected genes encoding collagens in the fibroblast clusters ($n=315$, 266, 219, 195, 175 and 155 fibroblasts for clusters 1, 2, 4, 5, 6 and 7, respectively). **e**, Differences in pathway activities scored per cell by GSVA between the different fibroblast clusters. Shown are t values from a linear model, corrected for patient of origin. **f**, Heatmap of the t values of AUC scores of expression regulation by transcription factors, as estimated using SCENIC, per fibroblast cluster. Shown are t values from a linear model for difference between cells from one cluster and cells from all other clusters, corrected for patient of origin, and this for all transcription factors having at least one t value exceeding 6. **g**, tSNE plots of fibroblasts, color-coded for (left) the expression of HOXB2 and MEF2C (top and bottom, respectively), for the AUC of the estimated regulon activity of these transcription factors (middle) and for the GSVA estimates of the indicated pathways (right).

(Supplementary Fig. 5) and bulk RNA-seq from TCGA (Fig. 4g). Despite this diverseness, tSNE plots showed a rather poor separation of clusters, suggesting that they represent diverse cell states on a graded scale rather than separate entities, in line with the spectrum model of macrophage activity[30]. Accordingly, marker gene analysis for clusters 1, 2 and 3 failed to identify specifically expressed genes (Supplementary Table 3).

Strikingly, tSNE plotting revealed a dichotomy between tumor- and non-malignant lung-derived macrophages (Fig. 4e), in line with a recent report characterizing macrophages from a single patient[7]. When contrasting pathway expression levels in both subsets (Fig. 4h), we noticed a strong reduction of inflammatory response, TNF-α-induced proliferation and reactive oxygen species producing pathways in tumor-derived macrophages, which are hallmarks of the M2-like, pro-tumoral subtype of macrophages described in murine cancer models[31]. SCENIC revealed that genes regulated by the IRF2, IRF7, IRF9 and STAT2 transcription factors were upregulated in a subset of tumor-associated macrophages, whereas genes decreased in expression were regulated by Fos/Jun and IRF8 (Fig. 4i,j). Notably, IRF2 has immunosuppressive roles in macrophages, while Fos/Jun can enhance inflammatory responses of macrophages and IRF8 favors M1 polarization (Fig. 4j)[32,33]. These data support M2 polarization of tumor macrophages in human tumors and identify compelling candidate transcription factors underlying these changes in NSCLC.

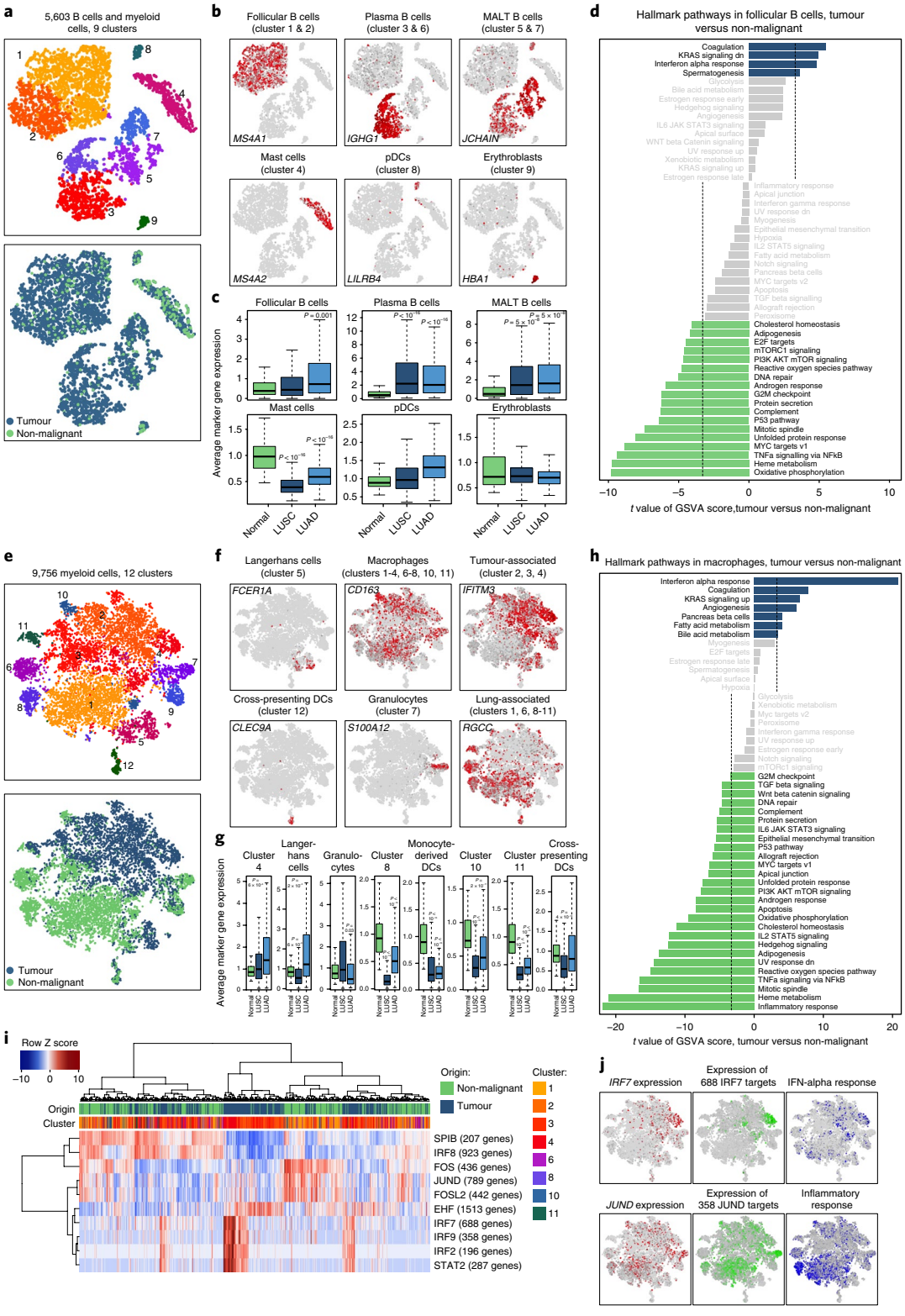
Tumor T-cell transcriptomes suggest novel immunotherapy targets. With 24,911 cells detected, T cells represent the most prevalent cell type. Reclustering revealed nine clusters, which were designated as regulatory T cells (*FOXP3*+, cluster 7), natural killer and natural killer T cells (*FGFBP2*+, cluster 6), CD8+ T cells (*CD8A*+; 2, 4, 5 and 8) and CD4+ T cells (*CD4*+; 1, 3 and 9) (Fig. 5a,b; Supplementary Fig. 13). In cluster 2, we also detected minor populations of innate lymphoid type 1-like cells and γδ T cells (Supplementary Fig. 13). Cells from all clusters were detected in all patients, apart from cluster 4 which derived predominantly from one tumor. CD8+ and regulatory T cells appeared enriched in the tumor and CD4+ T cells and natural killer cells appeared depleted, with the exception of CD4+ cluster 9, which was only present in tumors. These differences were confirmed in a recent study of the lung tumor immune landscape by CyTOF mass cytometry (Supplementary Fig. 14)[7], in 40,250 additional cells (Supplementary Fig. 5) and in TCGA data (Fig. 5c). The enrichment of T-cell clusters in either tumor or non-malignant lung samples suggests a strong influence of the tumor on the T-cell transcriptome. We therefore compared pathway expression levels between tumor and non-malignant lung-derived T cells. This revealed pervasive changes, mostly coherent between T-cell types (Fig. 5d), including an increased glycolysis and a decreased oxidative phosphorylation. Also, cell proliferation pathways were generally low in tumor-derived T cells, except within the tumor-specific, proliferative CD8+ T-cell cluster 8.

However, some pathways were differentially regulated between T-cell subtypes. For CD8+ T cells, we noticed a highly proliferative cluster (mitosis, G2M checkpoint, E2F targets; cluster 8) and two clusters with high allograft rejection activities, which in this context probably relate to cells showing higher reactivity to cancer-cell-encoded neo-epitopes (clusters 4 and 8) (Fig. 5e). These two clusters also show strong IFN-γ and IFN-α responses, higher rates of transcription, Myc activity and high granzyme expression, all indicative of T-cell activation. In parallel, they express higher levels of immune checkpoint molecules, including approved targets PDCD1 and CTLA4, but also others currently targeted in clinical trials (LAG3, TIGIT, HAVCR2/TIM3, CD27 and TNFRSF9/CD137) (Fig. 5f)[34]. All of these molecules also correlated with T-cell activity as measured by mean granzyme (*GZMA*, *GZMB* and *GZMH*) expression (Fig. 5g), indicating that higher cytotoxic activities are curtailed by high checkpoint expression. We identified several other molecules exhibiting a similarly strong correlation with granzyme expression, suggesting they could represent novel appealing checkpoint molecules (Fig. 5g). Other pathways correlating to high allograft rejection activity in CD8+ T cells include increased oxidative phosphorylation and fatty acid oxidation (Fig. 5e), suggesting that targeting these pathways may enhance immunotherapy. A recent study indeed indicated that promoting fatty acid catabolism in CD8+ tumor-infiltrating T lymphocytes enhances their ability to slow tumor progression[35].

Alveolar and epithelial cells. Alveoli are composed of several cell types and make up an important fraction of normal lungs. Previously, 198 alveolar cells from mice were characterized by scRNA-seq[36]. The 1,710 alveolar cells detected here recapitulate many of the markers described in murine lungs but extend them into human biology, in number of cell types and marker genes. Indeed, we not only detected flat alveolar type 1 (AT1) cells (*AGER, CAV1*; cluster 3), surfactant-secreting cuboidal alveolar type 2 (AT2) cells (*SFTPC, ABCA3*; cluster 1) and secretory club cells (*SCGB1A1*; cluster 6), but also basal cells (*KRT15*; cluster 7), previously undetected[36] (Supplementary Fig. 15a,b).

Fig. 4 | B-cell and myeloid-like cell clusters in lungs and lung tumors. a, tSNE plot of 5,603 B-cell-like cells, color-coded by their associated cluster (top) or the sample type of origin (bottom). **b**, tSNE plot, color-coded for expression (gray to red) of marker genes for the cell types as indicated. pDCs, plasmacytoid dendritic cells. **c**, Average expression of marker genes for B-cell-like cells from each cluster (Supplementary Table 3) in TCGA samples from lung ($n=108$), LUSC ($n=501$) or LUAD ($n=513$). Expression of each gene is normalized to its average expression in non-malignant lung samples. Box plot center, box and whiskers correspond to median, IQR and 1.5×IQR, respectively. Data were analyzed using one-way ANOVA with Tukey's multiple-comparisons test. **d**, Differences in pathway activities scored per cell by GSVA, between follicular B cells (clusters 1 and 2) isolated from lung or lung tumors ($n=205$ and 2,470 follicular B cells from 5 patients). *T* values are from a linear model, corrected for effects from the patient of origin. Myeloid-like cell clusters in lungs and lung tumors. **e**, tSNE plot of 9,756 myeloid-like cells, color-coded by their associated cluster (top) or the sample type of origin (bottom). **f**, tSNE plot, color-coded for expression (gray to red) of marker genes for the cell types as indicated. Additional marker genes for cell types are shown in Supplementary Fig. 12. DCs, dendritic cells. **g**, Average expression of marker genes for myeloid-like cells from each cluster (Supplementary Table 3) in TCGA samples from lung ($n=108$), LUSC ($n=501$) or LUAD ($n=513$). Expression of each gene is normalized to its average expression in non-malignant lung samples. Analysis for clusters 1, 2, 3 and 6 failed to yield specific marker genes. Expression of each gene is normalized to its average expression in non-malignant lung samples. Box plot center, box and whiskers correspond to median, IQR and 1.5×IQR, respectively. Data were analyzed using one-way ANOVA with Tukey's multiple-comparisons test. **h**, Differences in pathway activities scored per cell using GSVA, between macrophages isolated from lung or lung tumors ($n=3,873$ or 4,201 macrophages, respectively). *T* values are from linear models, corrected for effects from the patient of origin. **i**, Heatmap of the AUC scores of expression regulation by transcription factors (regulon activity), as estimated using SCENIC, for each of the 8,074 macrophages. Shown are the five transcription factors having the highest difference in expression regulation estimates between tumor and non-malignant lung-derived macrophages, and vice versa five transcription factors having the highest difference between non-malignant lung-derived and tumor macrophages. **j**, tSNE plots of macrophages, color-coded for (top) the expression of (left and right) IRF9 and JUND, for the AUC of the estimated regulon activity of these transcription factors and for activities of pathways containing a subset of the target genes of these transcription factors, as estimated using GSVA.

In line with basal cells giving rise to squamous tumors, basal cell markers were highly upregulated in LUSC but not LUAD (Supplementary Fig. 15c). Finally, we detected cells expressing marker genes for COPD-induced injury (*MMP7*, *CXCL14*, *GDF15*; cluster 4)[22,37] and respiratory epithelial cells (*CYP4B1*; cluster 5; Supplementary Fig. 15b). As expected, AT1, AT2 and respiratory epithelial cells were almost exclusively found in non-malignant lung, a finding confirmed in TCGA and the 40,250-cell validation dataset (Supplementary Fig. 15c and Supplementary Fig. 5, respectively). Surprisingly, however, club cells appeared often tumor-derived, and club cell markers were also in TCGA more prevalent in lung tumors (Supplementary Fig. 15c). Whether and how club cells contribute to lung tumor biology remains to be elucidated.



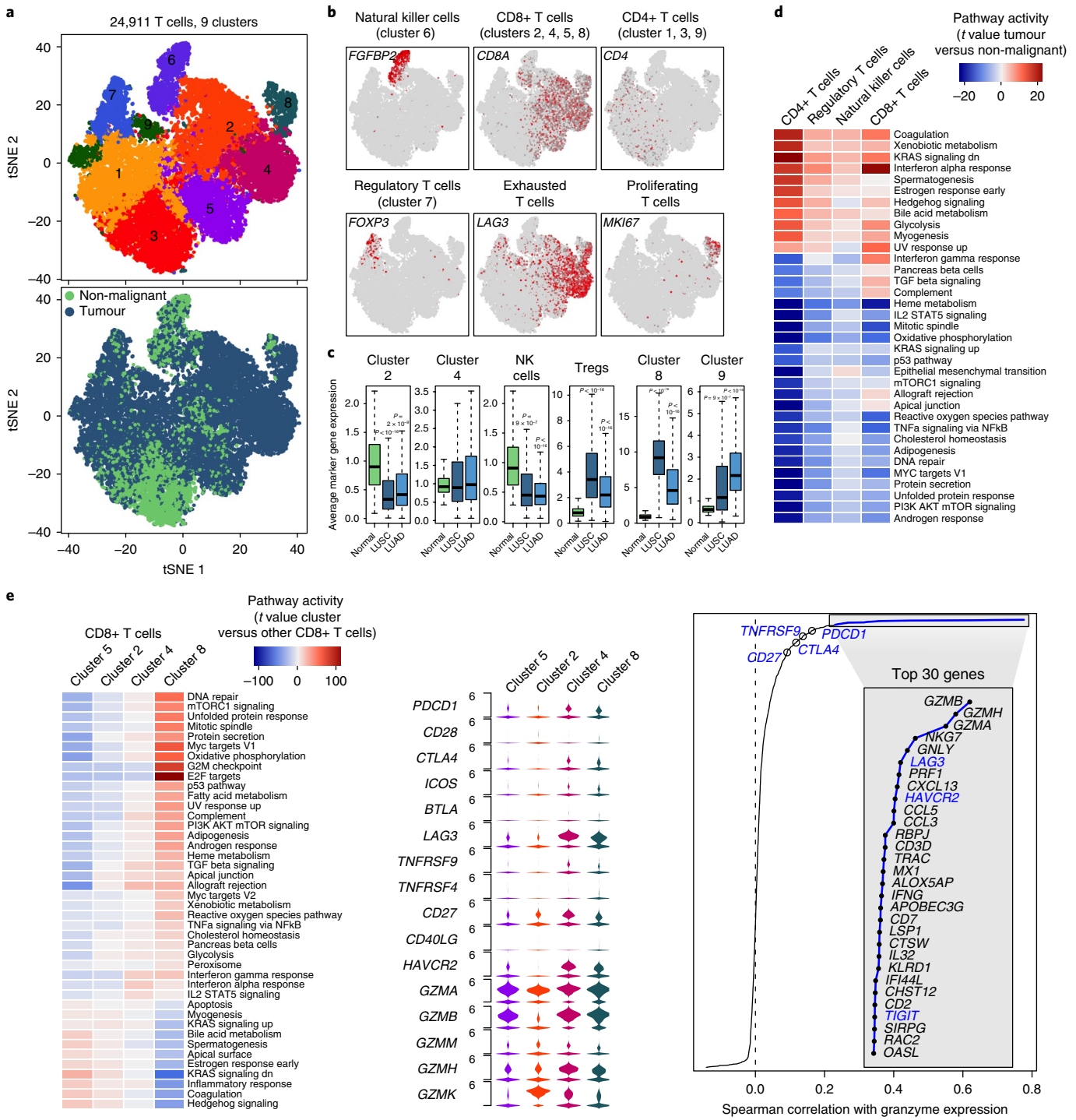


Fig. 5 | T-cell clusters in lungs and lung tumors. a, tSNE plot of 24,911 T cells, color-coded by their associated cluster (top) or the sample type of origin (bottom). **b**, tSNE plot, color-coded for expression (gray to red) of marker genes for the cell types, as indicated. **c**, Average expression of marker genes for T cells from each cluster (Supplementary Table 3) in TCGA samples from lung ($n=108$), LUSC ($n=501$) or LUAD ($n=513$). Expression of each gene is normalized to its average expression in non-malignant lung samples. Analysis for clusters 1, 3 and 5 failed to yield specific marker gene sets. Box plot center, box and whiskers correspond to median, IQR and 1.5 × IQR, respectively. Data were analyzed using one-way ANOVA with Tukey's multiple-comparisons test. **d**, Differences in pathway activities scored per cell using GSVA, between T cells isolated from lung or lung tumors. *t* values are independent of effects from the patient of origin. $n=3{,}547$ and 6,070 CD4 T cells, 1,091 and 10,949 CD8 T cells, 1,002 and 739 natural killer (T) cells and 226 and 1,287 regulatory T cells, respectively, derived from non-malignant and tumor tissue of 5 patients. **e**, As in **d**, but for CD8 T-cell clusters 5, 2, 4 and 8 ($n=$ respectively 421 and 2,704, 549 and 4,106, 10 and 3,244 and 111 and 895 CD8 T cells, derived from non-malignant and tumor tissue of 5 patients). **f**, Violin plots showing the smoothened expression distribution of selected genes involved in T-cell activity and in immune checkpoints, stratified by CD8 T-cell cluster ($n=3{,}125$, 4,655, 3,254 and 1,006 CD8 T cells of clusters 5, 2, 4 and 8, respectively, derived from 5 patients). **g**, Spearman correlation between activity of CD8 T cells ($n=8{,}915$), as measured by the average granzyme expression (*GZMA*, *GZMB* and *GZMH*), and expression of 1,704 CD8 T-cell specific genes (>3-fold overexpressed versus all other cells). Highlighted in blue are genes encoding known immune checkpoint molecules. NK, natural killer.

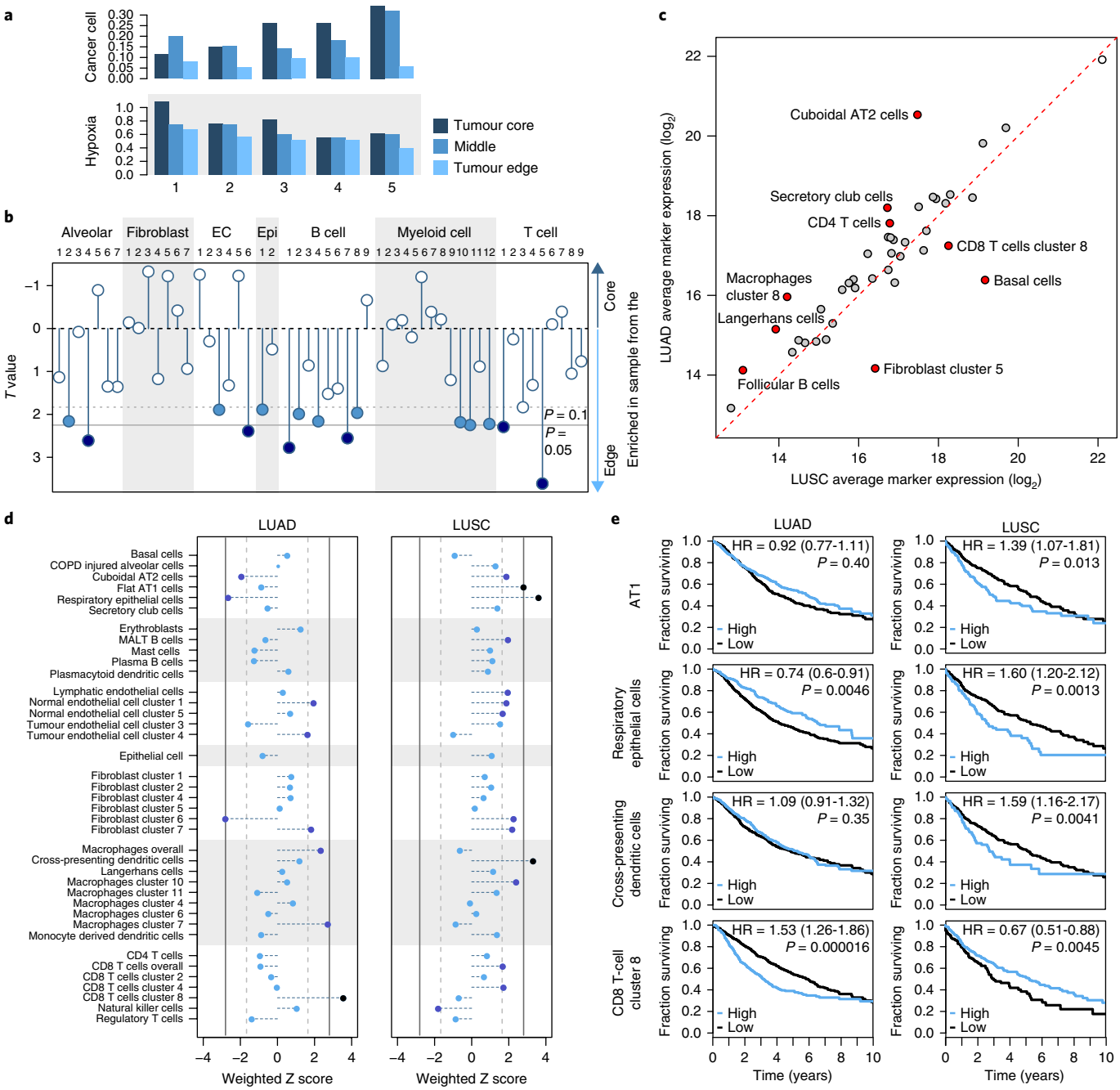


Fig. 6 | Distribution of stromal cells in tumor samples and their role as markers of patient survival. a, Cancer cells fraction and hypoxia marker gene expression for each tumor-derived sample. The five tumors included in this study were separated into three pieces, and each of them was marked as originating from either the core, the edge or the middle, in between edge and core. **b**, *t* values from a linear model, showing enrichment of stromal cell clusters in the tumor core or edge. The number of cells in each linear model is specified in Supplementary Table 3. **c**, Average expression of marker genes in LUAD ($n=501$) or LUSC ($n=513$) samples, characterized using RNA-seq in TCGA. Stromal cell types showing >2-fold change in expression between both cancer types are named and highlighted. **d**, Association between marker gene expression (continuous) in 1,027 LUAD samples (left) or in 545 LUSC samples (right) and patient survival. Weighted Z scores are denoted with points and a horizontal dashed line starting from 0, and are calculated using linear regression, corrected for age, gender and tumor stage. The vertical gray lines indicate threshold for nominal (light gray) and multiple-testing-corrected (dark gray) significance (respectively $P<0.05$ and false discovery rate <0.05). **e**, Kaplan–Meier survival curves for patients with LUAD or LUSC ($n=1,027$ or 545 patients, respectively), stratified for the average expression (binary: high versus low) of stromal cell marker genes as annotated in Supplementary Table 3. Stratification for high versus low expression was optimized as described[42]. Hazard ratios (HRs), with their 95% confidence intervals in brackets, and Cox regression *P* values obtained after correction for age, gender, study and stage are shown for four subclusters, respectively, for LUAD and LUSC.

Specific stromal cell subtypes are enriched at the more normoxic tumor edge. Hypoxia is associated with important aspects of tumor behavior, including angiogenesis, metastasis, metabolic reprogramming and immune escape. We therefore assessed whether certain cell subtypes reside more often in hypoxic areas of the tumor. Hypoxia levels were estimated for each of the three

samples obtained from the core-most, the edge-most and the intermediate fraction of the tumor, using a hypoxia metagene signature[38]. To avoid confounders, the signature was only applied to cancer cells. As expected, tumor core samples had more cancer cells, which were more hypoxic than samples from the edge, with midway samples showing intermediate levels ($P = 0.0051$ for cancer cells and 0.0018 for hypoxia; Fig. 6a). Increased hypoxia in the tumor core was confirmed by immunofluorescence for HIF1α and the hypoxia marker CA9 (Supplementary Fig. 16). Most stromal cell subtypes showed enrichment towards the more normoxic tumor edge (Fig. 6b). We observed, for instance, that CD4+ T-cell cluster 1 and CD8+ T-cell cluster 5, and follicular and MALT B cells, were enriched at the tumor edge. When subsequently assessing the influence of tumor hypoxia on marker gene expression (Supplementary Table 3) in 1,014 TCGA lung tumors, very similar patterns were observed, and this was for both LUAD and LUSC (Supplementary Fig. 17a). Some additional cell subtypes were also increased in hypoxic LUAD and LUSC tumors, including fibroblast cluster 5 and CD8+ T-cell cluster 8.

Correlation between stromal cell subtypes and tumor characteristics. Next, we correlated marker gene expression with tumor characteristics, such as histology and stage. While 33 of 42 cell subtypes for which marker gene expression signatures were available failed to show differences between LUAD and LUSC, 9 cell subtypes were different (Fig. 6c). For two of these nine cell subtypes, expression discrepancies probably reflect the cell of origin of lung tumors, with LUAD deriving from AT2 cells and LUSC from basal cells. For the remaining seven cell subtypes, such as CD8 T-cell cluster 8 and fibroblast cluster 5, correlations may reflect differences underlying histopathology. With respect to stage, squamous tumors showed a decrease in many stromal components at higher stages, whereas few stage-associated changes were observed in adenocarcinomas (Supplementary Fig. 17b).

Given the correlation between mutational load and tumor antigenicity, we correlated marker gene expression also with mutational load. This revealed that markers of CD8 T-cell cluster 8 were positively correlated with mutational load (Supplementary Fig. 18), whereas nearly all other stromal cell markers were associated with reduced mutational load. This arguably reflects an artefact of mutation calling: increases in the fraction of non-cancer cells will decrease the allelic frequency of somatic mutations, rendering them more difficult to detect. A similar trend was observed when assessing association to *TP53* or *RAS* (*KRAS*, *HRAS*, *BRAF*) mutations. Interestingly, *EGFR* mutations in LUAD were negatively associated with the CD8+ T-cell cluster 2, but positively with fibroblast cluster 6. A subpopulation of CD200+ cancer-associated fibroblasts was recently shown to enhance the apoptotic effects of gefitinib in *EGFR*-mutant lung cancers.[39] Notably, normal fibroblasts that belong to cluster 6 express lower levels of CD200 than those in clusters 1, 4, 5 and 7, suggesting cluster 6 fibroblasts to confer a selective growth advantage to *EGFR*-mutant lung tumors.

High stromal marker expression and decreased survival in LUSC, but not LUAD. Finally, we evaluated whether the relative presence of these cell subtypes impacts patient survival. We assessed, in addition to TCGA data, two other lung tumor data sets[40,41], such that together 1,027 LUAD and 545 LUSC samples were evaluated. Remarkably, we observed a consistent association between increased stromal marker gene expression and decreased survival in LUSC, but not LUAD, and this in a multivariate analysis corrected for age, gender and tumor stage (Fig. 6d). Furthermore, this association was not only evident in the pooled cohort (Fig. 3d), but also in each of the separate cohorts (Supplementary Fig. 19). Several cell subtypes were also associated with poor outcome in these cohorts, and for four stromal cell subtypes this survived multiple testing correction in the pooled analysis (false discovery rate <0.05 for LUSC: flat AT1 cells, respiratory epithelial cells and cross-presenting dendritic cells, and for LUAD: CD8+ T-cells cluster 8; Fig. 6d,e). Interestingly, several cell subtypes also showed opposite associations with survival when comparing LUAD versus LUSC (for example, CD8+ T-cell cluster 8, cuboidal AT2 cells, respiratory epithelial cells and fibroblast cluster 6), while other cell subtypes showed similar correlations (for example, endothelial cell cluster 1 and fibroblast cluster 7).

Discussion

Here, we present a comprehensive catalog of stromal cells in human lung tumors and non-malignant lung tissue at single-cell resolution. In describing key molecular differences between stromal cells co-opted by tumors and those in matching non-malignant samples, our analyses confirm many important observations made previously either in vitro, in bulk or using animal models, and highlight key areas for further advances in stromal cell biology. By identifying novel cell subtypes and altered pathways, by highlighting the cellular sources of stromal signals and by cataloging marker genes, this dataset will fuel advances in lung cancer diagnosis and therapy. It can serve to validate a priori hypotheses from independent data, but also highlight novel targets meriting functional validation.

While all cell types, subtypes and phenotypes cannot possibly be described here in full, some key observations emerge. Firstly, the lung TME is more complex and heterogeneous than hitherto appreciated. While previous scRNA-seq studies on tumors[1,12] analyzed far fewer stromal cells only derived from tumors, and clustered cells into major cell types only, we here analyzed 92,948 cells (84,341 stromal cells). This identified 52 stromal subtypes, including different types of tumor-associated fibroblasts, endothelial cells and tumor-infiltrating immune cells hitherto considered homogeneous. Each subtype showed divergent pathway activities, both between non-malignant and tumor tissue-resident counterparts, and between each other, suggesting that they represent distinct biological entities. Immunofluorescence confirmed the existence of four endothelial and fibroblast subtypes, and single-cell CyTOF data[7] of some immune subtypes. However, whether subtypes represent separate cell types or rather cell states acquired in response to TME stimuli is unclear, and at least in part a matter of semantics[42]. Lineage tracing represents an appealing avenue to elucidate this further.

Secondly, most clusters were composed of cells originating from different patients, while independent analysis of 3 additional patients validated 45 of 52 cell subtypes. However, marker gene expression in TCGA indicated that some cell subtype abundances differed between LUSC or LUAD, that they were influenced by tumor characteristics such as tumor stage and that they correlated with patient survival in LUSC but not LUAD or vice versa. Hence, while most stromal subtypes were detected in several tumors, their abundances and functions could differ between tumors. Intriguing questions remain as to whether these stromal cell phenotypes also exist in tumors affecting other organs, and whether they recur in metastases of lung tumors to other organs.

A third observation relates to the association of cell subtypes with patient survival: squamous tumors showed a negative correlation between many stromal cell markers and survival, independently of tumor stage. Whether this difference reflects a genuine effect of stromal cells on squamous cancer cell biology remains to be elucidated. Indeed, the stromal cell number in tumor samples could also reflect invasive properties of cancer cells, with more invasive tumors containing more stromal cells. Nevertheless, observations that some marker genes correlated robustly with survival in various cohorts suggest exciting opportunities for use of these subtype-specific marker genes as biomarkers for prognosis

but also for therapy response prediction. Most notably, low expression of CD8+ T-cell cluster 8 marker genes associated with improved survival in LUAD, but worse survival in LUSC. This cluster showed high granzyme and IFN expression, suggesting it represents CD8+ cytotoxic T cells. Indeed, cluster 8 also correlated positively with mutational load, was more frequent in hypoxic tumors and was characterized by high T-cell exhaustion marker expression (LAG3), in line with reports implicating tumor hypoxia in T-cell exhaustion[43].

Lastly, gene expression changes in tumor stroma suggest directions for the design of therapies. For instance, tumor endothelial cells downregulate immune cell homing pathways, while tumor CD8+ T cells upregulate fatty acid oxidation pathways. Additionally, analysis of CD8+ T-cell activity highlights established immune checkpoints, but also reveals several other coregulated molecules as potential novel immunotherapy targets. Likewise, SCENIC in macrophages predicts transcription factors responsible for the switch between antitumoral M1 and pro-tumoral M2 phenotypes. By targeting these, the macrophage transcriptome could thus rewire towards an M1 phenotype, with potential therapeutic benefit. Distinctive features of tumor stroma may thus represent vulnerabilities and provide exciting entry points for the design of novel therapies.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41591-018-0096-5.

Received: 16 October 2017; Accepted: 16 May 2018;
Published online: 9 July 2018

References

1. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
2. Albini, A. & Sporn, M. B. The tumour microenvironment as a target for chemoprevention. *Nat. Rev. Cancer* **7**, 139 (2007).
3. Vaupel, P., Kallinowski, F. & Okunieff, P. Blood flow, oxygen and nutrient supply, and metabolic microenvironment of human tumors: A review. *Cancer Res.* **49**, 6449–6465 (1989).
4. Eberhard, A. et al. Heterogeneity of angiogenesis and blood vessel maturation in human tumors: Implications for antiangiogenic tumor therapies. *Cancer Res.* **60**, 1388–1393 (2000).
5. Gordon, S. & Taylor, P. R. Monocyte and macrophage heterogeneity. *Nat. Rev. Immunol.* **5**, 953 (2005).
6. Sugimoto, H., Mundel, T. M., Kieran, M. W. & Kalluri, R. Identification of fibroblast heterogeneity in the tumor microenvironment. *Cancer Biol. Ther.* **5**, 1640–1646 (2006).
7. Lavin, Y. et al. Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses. *Cell* **169**, 750–765.e717 (2017).
8. Rittmeyer, A. et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): A phase 3, open-label, multicentre randomised controlled trial. *Lancet* **389**, 255–265 (2017).
9. Reck, M. et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N. Engl. J. Med.* **2016**, 1823–1833 (2016).
10. Reck, M. et al. Docetaxel plus nintedanib versus docetaxel plus placebo in patients with previously treated non-small-cell lung cancer (LUME-Lung 1): A phase 3, double-blind, randomised controlled trial. *Lancet Oncol.* **15**, 143–155 (2014).
11. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
12. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
13. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
14. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
15. Mazzone, M. et al. Heterozygous deficiency of PHD2 restores tumor oxygenation and inhibits metastasis via endothelial normalization. *Cell* **136**, 839–851 (2009).
16. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
17. Lin, C. Y. et al. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).
18. Baudino, T. A. et al. c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression. *Genes Dev.* **16**, 2530–2543 (2002).
19. Cantelmo, A. R. et al. Inhibition of the glycolytic activator PFKFB3 in endothelium induces tumor vessel normalization, impairs metastasis, and improves chemotherapy. *Cancer Cell* **30**, 968–985 (2016).
20. Arany, Z. et al. HIF-independent regulation of VEGF and angiogenesis by the transcriptional coactivator PGC-1alpha. *Nature* **451**, 1008–1012 (2008).
21. De Bock, K. et al. Role of PFKFB3-driven glycolysis in vessel sprouting. *Cell* **154**, 651–663 (2013).
22. Kambayashi, T. & Laufer, T. M. Atypical MHC class II-expressing antigen-presenting cells: Can anything replace a dendritic cell? *Nat. Rev. Immunol.* **14**, 719–730 (2014).
23. Tian, L. et al. Mutual regulation of tumour vessel normalization and immunostimulatory reprogramming. *Nature* **544**, 250–254 (2017).
24. Aibar, S. et al. SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083 (2017).
25. Wang, N. et al. Adenovirus-mediated overexpression of c-Jun and c-Fos induces intercellular adhesion molecule-1 and monocyte chemoattractant protein-1 in human endothelial cells. *Arterioscler. Thromb. Vasc. Biol.* **19**, 2078–2084 (1999).
26. Kalluri, R. The biology and function of fibroblasts in cancer. *Nat. Rev. Cancer* **16**, 582–598 (2016).
27. Gelse, K., Poschl, E. & Aigner, T. Collagens—Structure, function, and biosynthesis. *Adv. Drug Deliv. Rev.* **55**, 1531–1546 (2003).
28. Lin, Q., Schwarz, J., Bucana, C. & Olson, E. N. Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. *Science* **276**, 1404–1407 (1997).
29. Lu, J., Webb, R., Richardson, J. A. & Olson, E. N. MyoR: A muscle-restricted basic helix-loop-helix transcription factor that antagonizes the actions of MyoD. *Proc. Natl. Acad. Sci. USA* **96**, 552–557 (1999).
30. Xue, J. et al. Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity* **40**, 274–288 (2014).
31. Biswas, S. K. et al. A distinct and unique transcriptional program expressed by tumor-associated macrophages (defective NF-kappaB and enhanced IRF-3/STAT1 activation). *Blood* **107**, 2112–2122 (2006).
32. Gunthner, R. & Anders, H. J. Interferon-regulatory factors determine macrophage phenotype polarization. *Mediat. Inflamm.* **2013**, 731023 (2013).
33. Medzhitov, R. & Horng, T. Transcriptional control of the inflammatory response. *Nat. Rev. Immunol.* **9**, 692 (2009).
34. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252 (2012).
35. Zhang, Y. et al. Enhancing CD8+ T cell fatty acid catabolism within a metabolically challenging tumor microenvironment increases the efficacy of melanoma immunotherapy. *Cancer Cell* **32**, 377–391.e39 (2017).
36. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
37. Shaykhiev, R. et al. Smoking-induced CXCL14 expression in the human airway epithelium links chronic obstructive pulmonary disease to lung cancer. *Am. J. Respir. Cell. Mol. Biol.* **49**, 418–425 (2013).
38. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* **102**, 428–435 (2010).
39. Ishibashi, M. et al. CD200-positive cancer associated fibroblasts augment the sensitivity of Epidermal Growth Factor Receptor mutation-positive lung adenocarcinomas to EGFR Tyrosine kinase inhibitors. *Sci. Rep.* **7**, 46662 (2017).
40. Djureinovic, D. et al. Profiling cancer testis antigens in non-small-cell lung cancer. *JCI Insight* **1**, e86837 (2016).
41. Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma. et al. Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat. Med.* **14**, 822–827 (2008).
42. Clevers, H. et al. What is your conceptual definition of 'cell type' in the context of a mature organism? *Cell Syst.* **4**, 255–259 (2017).
43. Zhang, Y. & Ertl, H. C. Starved and asphyxiated: How can CD8+ T cells within a tumor microenvironment prevent tumor progression. *Front. Immunol.* **7**, 32 (2016).

Acknowledgements

We thank M. De Waegeneer, T. Van Brussel, G. Peuteman, E. Vanderheyden and B. Tembuyser for technical assistance. This work was supported by a VIB TechWatch Grant to D.L. and B.T., Foundation Against Cancer grants to S.Aerts (2016-070) and E.W., ERC Consolidator Grants to S.Aerts (724226_cis-CONTROL) and D.L. (CHAMELEON), Funds for Research - Flanders grants to H.D. (1701018N) and D.L.

(G065615N), an Austrian Science Fund (FWF) grant to A.P. (J3730-B26) and KU Leuven grants to D.L. and S.Aerts (PFV/10/016 SymBioSys), and to B.T. (BOFZAP).

Author contributions

D.L. and B.T. designed and supervised the study and wrote the manuscript. E.W. supervised sample collection and clinical annotation, with important help from H.D., A.P., K.V.d.E., B.W., E.V., P.D.L. and J.V. B.T. performed data analysis, with significant contributions from B.B., S.Ai., S.Ae. and A.B. D.N. and B.T. performed immunohistofluorescence analyses. O.B., A.L., P.C. and S.Ae. contributed critical data interpretation. All of the authors have read or provided comments on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0096-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.L. or B.T.

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

Methods

Patients. This study was approved by the local ethics committee at the University Hospital Leuven (B322201422081) and we complied with all relevant ethical regulations. Only patients with untreated, primary, non-metastatic lung tumors that underwent lung lobe resection with curative intent and that provided informed consent were included in this study.

Preparation of single-cell suspensions. Following resection in the operating room, samples from the tumor and adjacent non-malignant lung tissue from the same resection specimen at maximal distance (>5 cm) from the tumor were isolated and transported rapidly to the research facility. On arrival, samples were rinsed with PBS and the tumor sample macroscopically examined for tumor positioning. The tumor sample was subsequently divided into three pieces, with one piece containing mainly tissue derived from the tumor core, one piece containing tissue mainly derived from the tumor edge and a third piece originating from the position intermediate to the other two samples. Each sample was subsequently minced on ice to smaller pieces of less than 1 mm^3 and transferred to 10 ml digestion medium containing 0.2% collagenase I/II (ThermoFisher Scientific), DNAse I (Sigma) and 25 units dispase (Invitrogen) in DMEM (ThermoFisher Scientific). Samples were incubated for 15 min at 37 °C, with manual shaking every 5 min. Samples were then vortexed for 10 s and pipetted up and down for 1 min using pipettes of descending sizes (25 ml, 10 ml and 5 ml). Next, 30 ml ice-cold PBS, pH 7.4, (ThermoFisher Scientific) containing 2% fetal bovine serum (ThermoFisher Scientific) was added and samples were filtered using a 40-µm nylon mesh (ThermoFisher Scientific). Following centrifugation at 120× *g* and 4 °C for 5 min, the supernatant was decanted and discarded, and the cell pellet was resuspended in 2 ml red blood cell lysis buffer and transferred to a 2-ml DNA low bind tube. Following a 5-min incubation at room temperature, samples were centrifuged (120×*g*, 4 °C, 5 min) using a swing-out rotor. Samples were next resuspended in 1 ml PBS containing 8 µl UltraPure BSA (50 mg ml^{-1}; AM2616, ThermoFisher Scientific) and filtered over Scienceware Flowmi 40-µm cell strainers (VWR) using wide-bore 1 ml low-retention filter tips (Mettler-Toledo). Next, 10 µl of this cell suspension was counted using an automated cell counter (Luna) to determine the concentration of live cells. Throughout the dissociation procedure, cells were maintained on ice whenever possible, and the entire procedure was completed in less than 1 h (typically ~45 min) to avoid dissociation-associated artefacts recently described[13]. By using a dissociation signature[13] to detect dissociation-associated changes in gene expression, a positive signal for less than 2% of cells was detected (Supplementary Fig. 2a).

Droplet-based scRNA-seq. 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), aiming for an estimated 4,000 cells per library and following the manufacturer's instructions. Samples were processed using kits pertaining to either the V1 or V2 barcoding chemistry of 10x Genomics (Supplementary Table 2). Single samples are always processed in a single well of a PCR plate, allowing all cells from a sample to be treated with the same master mix and in the same reaction vessel. For each patient, all samples (non-malignant and tumor) were processed in parallel in the same thermal cycler. Libraries were sequenced on an Illumina HiSeq4000, and mapped to the human genome (build hg19) using CellRanger (10x Genomics). Gene positions were annotated as per Ensembl build 85 and filtered for biotype (only protein-coding, long intergenic non-coding RNA, antisense, immunoglobulin or T-cell receptor).

Single-cell gene expression quantification and determination of the major cell types. Raw gene expression matrices generated per sample using CellRanger (version 2.0.0) were combined in R (version 3.3.2—*Sincere Pumpkin Patch*), and converted to a Seurat object using the Seurat R package (version 1.4.0.7)[11]. From this, all cells were removed that had either fewer than 201 UMIs, over 6,000 or below 101 expressed genes, or over 10% UMIs derived from mitochondrial genome. From the remaining 52,698 cells, gene expression matrices were normalized to total cellular read count and to mitochondrial read count using linear regression as implemented in Seurat's *RegressOut* function. As a result, none of the principle components subsequently identified were correlated with transcript count (data not shown). From the remaining 52,698 cells, variably expressed genes were selected as having a normalized expression between 0.125 and 3, and a quantile-normalized variance exceeding 0.5. To reduce dimensionality of this dataset, the resulting 2,192 variably expressed genes were summarized by principle component analysis, and the first 8 principle components further summarized using tSNE dimensionality reduction using the default settings of the *RunTSNE* function. Cell clusters in the resulting two-dimensional representation were annotated to known biological cell types using canonical marker genes (Supplementary Fig. 1). Of note, very few stromal cells (~2%) were positive for cell proliferation markers (Supplementary Fig. 4). We therefore opted not to correct our gene expression matrices for effects of cell cycle.

Subclustering of the major cell types. To identify subclusters within these eight cell types, we reanalyzed cells belonging to each of these eight cell types separately. Specifically, we applied dimensionality reduction using principle component analysis in each cell type on variably expressed genes as described above. To identify which principle components were informative, we applied Horn's parallel analysis for principle component analysis[44] as implemented in the R *paran* package (version 1.5.1.), selecting those principle components having eigenvalues that exceed the eigenvalues generated using ten random permutations by >50%. Using the graph-based clustering approach implemented in the *FindClusters* function of the Seurat package, with a conservative resolution of 0.5 and otherwise default parameters, each cell type was reclustered by its principle components. Notably, subclustering was robust to alterations in the number of principle components, in the resolution or in the *K* parameter (Supplementary Fig. 3a–c). Moreover, few of the subclusters identified contained many cells wherein less than 300 genes were detected, indicating that increasing the threshold of 100 genes will not affect our results (Supplementary Fig. 20). This yielded 64 subclusters (52 stromal subclusters) in total, as listed in Supplementary Table 3. For visualization purposes, these informative principle components were converted into tSNE plots as above.

Validation dataset and random forest mapping. To validate the presence of these stromal subclusters, we analyzed three additional patients (three tumor samples and one non-malignant lung sample) by scRNA-seq as described above. We found that 40,250 cells passed the quality control criteria described above, and these were used to generate gene expression matrices. To assign cells to one of the major cell types, we clustered them as described above, again enabling us to classify them as fibroblasts or endothelial, epithelial, alveolar, myeloid, B, T or cancer cells.

To assess to which of the cell subclusters these cells correspond, we generated a Random Forrest[45] classifier using the *ClassifyCells* function in Seurat. This assigns the 40,250 cells from the validation set of 3 patients to the 52 stromal cell subclusters identified in the discovery set of 5 patients.

Identification of marker genes. To identify marker genes for each of these 64 subclusters within these 8 cell types, we contrasted cells from that subcluster to all other cells of that subcluster using the Seurat *FindMarkers* function. Marker genes were required to have an average expression in that subcluster that was >2.5-fold higher than the average expression in the other subclusters from that cell type, and a detectable expression in >15% of all cells from that subcluster. Additionally, marker genes were required to have the highest mean expression in that subcluster, out of all 64 subclusters. This yielded a list of in total 402 marker genes (Supplementary Table 3) for 51 subclusters (42 stromal cell subclusters), whereas for 13 subclusters we failed to identify marker genes. When analyzing marker genes for several subclusters in aggregate, such as for tumor endothelial cells (endothelial cell clusters 3 and 4) or for macrophages (myeloid clusters 1–4, 6–8, 10 and 11), we simply combined the marker genes for all associated subclusters.

Correlation to TCGA data. To assess the role of stromal cells in a larger compendium of tumors, we assessed their expression in bulk RNA-seq data from TCGA. Specifically, we downloaded pre-processed gene expression data (fragments per kilobase per million fragments, upper quartile normalized) as well as clinical data for primary solid tumors and normal solid tissue, for LUAD (TCGA-LUAD) and LUSC (TCGA-LUSC), using the Bioconductor *TCGAbiolinks* package (version 2.2.10). To assess per cell type the combined expression of marker genes for each subcluster, we generated boxplots without outliers of the average expression of each marker gene, after log-normalizing the expression of each gene to an average expression of 1 in the normal lung samples. Other variables (age, gender, patient survival, tumor stage, tumor cell percentage, mutational burden) were extracted from the clinical data downloaded using *TCGAbiolinks*. Hypoxia marker gene expression was categorized to 1, 2 and 3 (normoxic, intermediate and hypoxic) as described[46]. To assess the correlation of each set of marker genes to other clinical variables, their individual expression was averaged per patient and per set of marker genes. To assess correlations, these values were converted to the corresponding rank per tumor subtype (LUAD or LUSC). These ranks were subsequently included in a generalized linear model using R, together with patient age and gender when correlating with tumor stage, and with age, gender and tumor stage when correlating with tumor hypoxia or mutational burden. To assess for correlations with survival, we applied a Cox proportional hazards model (implemented in the R *survival* package version 2.41-3) that included age, gender and tumor stage in addition to the mean expression of the marker genes. As a validation cohort, we downloaded clinical and gene expression data from 2 additional studies: one of 108 LUAD samples and 67 LUSC samples, described by Djureinovic and colleagues[40], and one of 443 LUAD samples, described by Shedden and colleagues[41]. We assessed effects of marker gene expression on patient survival in an identical manner as for TCGA. Z scores from the TCGA and the validation cohort were combined using the weighted Z method[47]. For Kaplan–Meier plots, marker gene expression categorization was optimized as described[42].

Gene set variation analysis (GSVA). Pathway analyses were predominantly performed on the 50 hallmark pathways described in the molecular signature database[16], exported using the *GSEABase* package (version 1.36.0). We also assessed metabolic pathway activities using a described curated dataset[48]. To reduce pathway overlaps and pathway redundancies, each gene set associated with a pathway was trimmed to only contain unique genes, and all genes associated to two

or more pathways were removed. Most gene sets retained >70% of their associated genes. Next, to assign pathway activity estimates to individual cells, we applied GSVA[49] using standard settings, as implemented in the *GSVA* package (version 1.22.4).

SCENIC analysis. The SCENIC analysis was run as described[24] on the 52,698 cells that passed the filtering, using the 20-thousand motifs database for RcisTarget and GRNboost (SCENIC version 0.1.5, which corresponds to RcisTarget 0.99.0 and AUCell 0.99.5; with RcisTarget.hg19.motifDatabases.20k). The input matrix was the normalized expression matrix, output from Seurat, from which 9,919 genes passed the filtering (sum of expression $>3\times0.005\times52{,}698$ and detected in at least 0.5% of the cells).

Clustering using alternative tools. Apart from Seurat, we tested SCENIC[24], SC3[50], Cidr[51] and RCA[52]. SC3, Cidr and RCA were run using default settings. For SCENIC, we applied *k*-means clustering on the regulon activity matrix. All tools were run on the six main cell types discussed in this study: fibroblasts and endothelial, alveolar, myeloid, T and B cells. We were unable to run RCA on T cells. We performed a pairwise comparison on the output clusters of each of these five methods using the normalized mutual information criterion. To test which method was most recurrently in best agreement with the other four methods, we tabulated, for each of the five methods, the method generating the highest normalized mutual information, and this for each of the six cell types.

Analysis of differential pathway or regulon activities. To assess differential activities of pathways (GSVA) or regulons (SCENIC) between sets of cells (for example, derived from tumor or normal samples, or belonging to different subclusters), we contrasted the activity scores for each cell using a generalized linear model. To avoid inflating signals because of interindividual differences (for example, in the relative frequencies of cells from different patients), we always included the patient of origin as a categorical variable. Results of these linear models were visualized using bar plots or heatmaps. For the latter, pathways or regulons that did not show significant changes (Benjamini–Hochberg-corrected *P* value >0.05) in any of the sets of cells contrasted in one analysis were not visualized.

Comparison to bulk RNA-seq. Bulk RNA-seq was performed as described[46], on a tumor sample adjacent to sample 1 and a non-malignant lung sample adjacent to sample 19. Read counts per gene were normalized to gene length and to the total read count, and directly compared to the sum of UMIs per gene for the corresponding single-cell sample. Differential expression analysis was performed as described[46], and a ranked list of genes upregulated and downregulated in expression analyzed for their ontology using GOrilla. Cell concentrations in the bulk RNA-seq were estimated by quadratic programming using the R quadprog package (version 1.5-5), using expression of the ten most differentially expressed marker genes per cell type as input and constraining the model by requiring the combined concentration of all cell types to be 1.

Immunohistofluorescence analysis. For immunohistofluorescence confirmation of fibroblast and endothelial cell subtypes, an independent set of NSCLC samples was selected (five LUAD and three LUSC NSCLC patients). For tumor-specific cell subtype markers, we stained sections from eight tumor samples and from two non-malignant lung samples. For non-malignant lung sample markers, we stained sections from eight tumor samples and eight non-malignant lung samples.

Histopathology and immunohistochemistry. Tissue samples from representative lesions were collected and fixed as described[53]. Sections of 5-µm thickness obtained from the paraffin-embedded tissues (Thermo Scientific Microm HM355S microtome) were mounted on Superfrost Plus Adhesion slides (Thermo Scientific) and routinely stained with hematoxylin and eosin (Diapath #C0302 and #C0362) for histopathological examination.

The following antibodies and dilutions were used for detecting the respective proteins: anti-ACKR1 (rabbit, 1:100, Sigma-Aldrich, hpa016421, lot number: R05967), anti-Carbonic Anhydrase IX (rabbit, 1:1,000, Novus, NB100-417, no lot number known), anti-CD31 (mouse, 1:50, Dako, M082301, lot number: 20049471, clone JC/70A), anti-Claudin-5 (rabbit, 1:2,000, Abcam, ab131259, lot number: GR236334-15, clone EPR7583), anti-COL1A (rabbit, 1:12,000, Abcam, ab138492, lot number: GR247379-37, clone EPR7785), anti-COX4I2 (rabbit, 1:500, LSBio, LS-B9672, lot number: 57931), anti-EDNRB (rabbit, 1:500, Sigma-Aldrich, hpa027546, lot number: R26734) and anti-Hif-1 alpha (rabbit, 1:1,000, Abcam, ab2185, lot number: GR310634-1).

Furthermore, the PerkinElmer Opal 4-Color Manual IHC Kit (PerkinElmer, NEL810001KT) was used for the tyramide signal amplification according to the manufacturer's protocol. For introduction of the secondary HRP the Envision+ /HRP goat anti-Rabbit (Dako Envision+ Single Reagents, HRP, Rabbit, Code K4003) was used for antibodies raised in rabbit. The various proteins were detected using the OPAL 520, OPAL 570 and OPAL 690 reagents, respectively. For anti-CD31 (mouse, 1:50, Dako, M082301), the protocol was adapted from the PerkinElmer Opal 4-Color Manual IHC Kit: the antibody incubation was done overnight at +4 °C and for introduction of the secondary-HRP the anti-mouse biotin (1:200, Jackson, 715-065-150) and Streptavidin-HRP Conjugate (1:100, PerkinElmer, NEL750001EA) were used before applying the OPAL reagents.

Microscope image acquisition and processing. Images were acquired on a Zeiss Axio Scan.Z1 using a ×20 objective and ZEN 2 software (Zeiss). Image processing was done using QuPath (version 0.1.2). Specifically, following visual inspection of the staining results, cells were first automatically detected using the DAPI channel (cell size constrained between 5 and 400 μm^2). Next, a random trees cell classifier was generated using QuPath. Specifically, for one slide out of all slides stained for one set of proteins, three or four sets of cells were selected: one set that was positive for the general cell type marker (CLDN5 or CD31 for endothelial cells, COL1A1 for fibroblasts), one set that was negative for that general cell type marker and one or two sets of cells positive for both the subtype-specific marker (ACKR1, FIGF, EDNRB or COX4I2) and the general cell-type marker (CLDN5, CD31 or COL1A1). Using these three or four sets of cells, a random trees classifier was generated. Cell classification was visually verified to have occurred correctly. Next, for each tumor or non-malignant lung section, a representative region was selected, containing at least 6,000 cells. On these cells, the random trees classifier was subsequently applied. This process was reiterated for all other tumor sections stained for the same set of markers. The resulting cell identities were then exported and processed in R. Specifically, for each set of 1,000 consecutive cells, cell frequencies were generated, which were summarized using boxplots.

To measure hypoxia, we manually annotated the tumor core and edge and performed automatic cell detection as described above. For these cells, mean nuclear and cytoplasmatic signals of HIF1A and CAIX, respectively, were calculated and compared between pairwise core and edge.

Statistics and reproducibility. No statistical method was used to predetermine sample sizes. For all experiments, samples from a single patient were processed in parallel, and cells for each sample of one patient were processed for scRNA-seq (10x Genomics) at the same time, but in separate lanes and vials.

Box plots were generated using the R base package and default parameters. Hence, the boxes span the interquartile range (IQR; from the 25th to the 75th percentiles), with the centerline corresponding to the median. Lower whiskers represent the data minimum or the 25th percentile minus 1.5×IQR, whichever is greater. Upper whiskers represent the data maximum or the 75th percentile plus 1.5×IQR (lower), whichever is lower.

Violin plots were generated using the beanplot R package, and data distribution band width was estimated by kernel density estimation, as per the built-in 'nrd0' option.

Bar plots indicate mean ± standard error of mean, and include individual data points.

Given the number of data points represented on box and violin plots, we opted not to display each data point, as this would obscure the overall distribution.

Comparisons between two groups were done using unpaired two-tailed *t*-tests. One-way analysis of variance (ANOVA) with Tukey's multiple comparisons tests were used for multiple group comparisons. Linear models were generated when multiple parameters were taken into account. Fitting of Cox proportional hazards regression models was done using the coxph function implemented in the R *survival* package, with tied death times handled using the Breslow method. All statistical analyses and presentation were performed using R.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. All raw sequencing data are available in ArrayExpress under accessions E-MTAB-6149 and E-MTAB-6653. Also, Rds files were uploaded. These can be imported in CellView to visualise clusters, scroll through tSNE projections and explore gene expression. Moreover, scRNA-seq source data were formatted as .loom files, which can be visualized in an interactive manner through SCope (https://gbiomed.kuleuven.be/scRNAseq-NSCLC)[54]. Finally, gene expression data for all 52 clusters are available in Supplementary Table 4, and cluster-specific gene expression data for tumor-derived and non-malignant lung-tissue-derived cells are available in Supplementary Table 5 (only for clusters having >100 cells from both sources).

References

44. Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–185 (1965).
45. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
46. Thienpont, B. et al. Tumour hypoxia causes DNA hypermethylation by reducing TET activity. *Nature* **537**, 63–68 (2016).
47. Whitlock, M. C. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18**, 1368–1373 (2005).
48. Gaude, E. & Frezza, C. Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. *Nat. Commun.* **7**, 13041 (2016).

49. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7 (2013).
50. Kiselev, V. Y. et al. SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483 (2017).
51. Lin, P., Troup, M. & Ho, J. W. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 59 (2017).
52. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708 (2017).
53. Wauters, E. et al. DNA methylation profiling of non-small cell lung cancer reveals a COPD-driven immune-related signature. *Thorax* **70**, 1113–1122 (2015).
54. Kristofer, D. et al. A single-cell transcriptome atlas of the aging Drosophila brain. *Cell* https://doi.org/10.1016/j.cell.2018.05.057 (2018).

nature research

Corresponding author(s): Diether Lambrechts, Bernard Thienpont

☐ Initial submission ☒ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Sample size was not determined a priori. About 3000 cells were analyzed per sample, as higher cell numbers would have yielded unwanted amounts of cell duplicates.

2. Data exclusions

Describe any data exclusions.

No samples were excluded from the analysis. Cells with UMI counts below 200 or gene counts below 100 were excluded because they represent a poor quality or empty droplets. Cells having over 10% mitochondrial reads were excluded as these likely represent apototic cells. Cells expressing over 6000 genes were excluded as they likely represent duplicates. Exclusion criteria were pre-established.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All attempts at replication were successful

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

All 8 patients had untreated, non-metastatic NSCLC , and underwent lung lobe resection with curative intent. Cells were allocated to cell types using established marker gene expression patterns. No other experimental groups were established.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not relevant for our study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Raw gene expression matrices generated per sample using CellRanger (version 2.0.0) were combined in R (version 3.3.2 - Sincere Pumpkin Patch), and converted to a Seurat object using the Seurat R package (version 1.4.0.7).To identify which PCs were informative, we applied Horn's parallel analysis for PC analysis42 as implemented in the R paran package (version 1.5.1.). TCGA data were downloaded using the Bioconductor TCGAbiolinks package (version 2.2.10). To assess for correlations with survival, we applied a Cox proportional hazards model (implemented in the R survival package version 2.41-3) which included age, gender and tumour stage in addition to the ranked expression of the marker genes. Pathway data were exported from the molecular signature database using the GSEABase package (version 1.36.0). Pathway activity estimates were obtained using the GSVA package (version 1.22.4). The SCENIC analysis was done using the 20-thousand motifs database for RcisTarget and GRNboost (SCENIC version 0.1.5, which corresponds to RcisTarget 0.99.0 and AUCell 0.99.5; with RcisTarget.hg19.motifDatabases.20k). Images were acquired on a Zeiss Axio Scan.Z1 using ZEN 2 software (Zeiss). Image processing was done using QuPath (version 0.1.2).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Following antibodies and dilutions were used for detecting the respective proteins: anti-ACKR1 (rabbit, 1:100, Sigma-Aldrich, hpa016421, lot number: R05967), anti-Carbonic Anhydrase IX (rabbit, 1:1000, Novus, NB100-417, no lot number known), anti-CD31 (mouse, 1:50, Dako, M082301, lot number: 20049471), anti-Claudin-5 (rabbit, 1:2000, Abcam, ab131259, lot number: GR236334-15, clone EPR7583), anti-COL1A (rabbit, 1:12000, Abcam, ab138492, lot number: GR247379-37, clone EPR7785), anti-COX4I2 (rabbit, 1:500, LSBio, LS-B9672, lot number: 57931), anti-EDNRB (rabbit, 1:500, Sigma-Aldrich, hpa027546, lot number: R26734), anti-Hif-1 alpha (rabbit, 1:1000, Abcam, ab2185, lot number: GR310634-1).
Antibodies were verified to stain specific cell subtypes as identified by scRNA-seq, and to costain with established markers of endothelial cells or fibroblasts. Anti-Carbonic Anhydrase IX (1), anti-CD31 (2), anti-COL1A (3) and anti-Hif-1 alpha (4) were moreover previously shown to stain their respective epitopes in immunohistochemistry of human tumours.

References:
1: Corbet et al. Interruption of lactate uptake by inhibiting mitochondrial pyruvate transport unravels direct antitumor and radiosensitizing effects. Nat Commun. 2018 9(1) p1208

2: Nikolić et al. Human embryonic lung epithelial tips are multipotent progenitors that can be expanded in vitro as long-term self-renewing organoids. Elife. 2017 6 e26575

3: Zhou et al. The prognostic value and pathobiological significance of Glasgow microenvironment score in gastric cancer. J Cancer Res Clin Oncol. 2017 143(5) p883

4: Doublier et al. HIF-1 activation induces doxorubicin resistance in MCF7 3-D spheroids via P-glycoprotein expression: a potential model of the chemo-resistance of invasive micropapillary carcinoma of the breast. BMC Cancer. 2012 12 p4

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

All relevant information was summarized in Supplementary table 1.