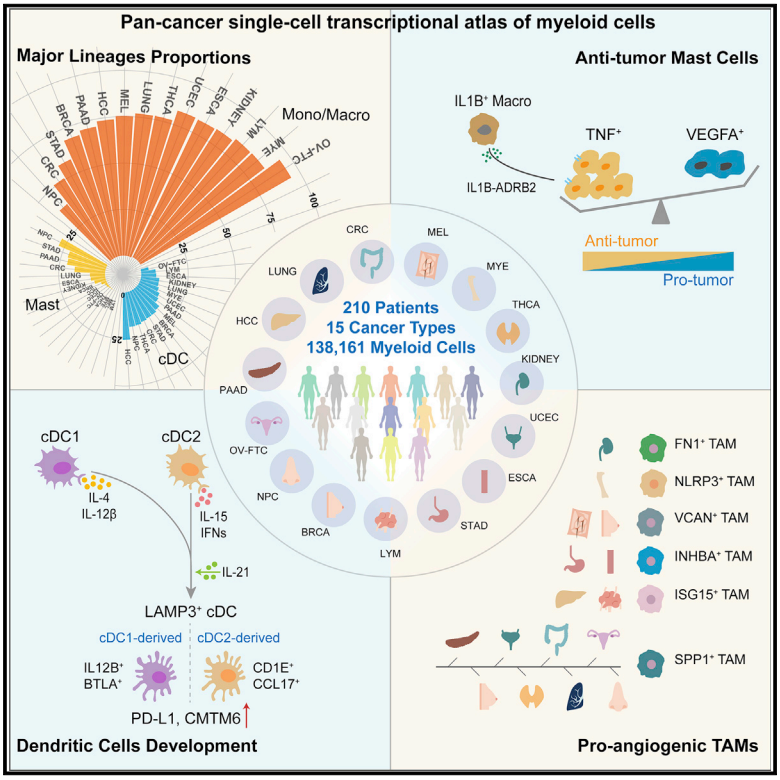
Cell

Resource

# A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells

## Graphical Abstract




## Highlights

- Pan-cancer analysis reveals heterogeneity in tumor-infiltrating myeloid cell composition
- The ratio of *TNF*$^+$ versus *VEGFA*$^+$ mast cells underlines their cancer-type-specific functions
- *LAMP3*$^+$ cDCs are widely present, with diverse developmental origins and functions
- Pro-angiogenic TAMs exhibit distinct expression profiles across different cancer types


## Authors

Sijin Cheng, Ziyi Li, Ranran Gao, ..., Zhaode Bu, Jiafu Ji, Zemin Zhang

## Correspondence

buzhaode@cjcrcn.org (Z.B.), jijiafu@hsc.pku.edu.cn (J.J.), zemin@pku.edu.cn (Z.Z.)



## In Brief

A systemic analysis of tumor-infiltrating myeloid cells across 15 human cancer types features the molecular basis of pro- versus anti-tumor effect of mast cells, the connection between developmental origins and functions in *LAMP3*$^+$ cDCs, and the diversity in pro-angiogenic tumor-associated macrophages across different cancers.


Cheng et al., 2021, Cell *184*, 792–809
February 4, 2021 
https://doi.org/10.1016/j.cell.2021.01.010

CellPress

CellPress
OPEN ACCESS

Cell

Resource

# A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells

Sijin Cheng,[1,13] Ziyi Li,[1,13] Ranran Gao,[1] Baocai Xing,[2] Yunong Gao,[3] Yu Yang,[1] Shishang Qin,[1] Lei Zhang,[1] Hanqiang Ouyang,[4] Peng Du,[5] Liang Jiang,[4] Bin Zhang,[6] Yue Yang,[7] Xiliang Wang,[1] Xianwen Ren,[1] Jin-Xin Bei,[8] Xueda Hu,[1] Zhaode Bu,[9,*] Jiafu Ji,[9,10,*] and Zemin Zhang[1,11,12,14,*]

[1]BIOPIC, Beijing Advanced Innovation Center for Genomics, School of Life Sciences, Peking University, Beijing 100871, China
[2]Department of Hepatopancreatobiliary Surgery I, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
[3]Department of Gynecologic Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
[4]Department of Orthopaedics, Peking University Third Hospital, Beijing Key Laboratory of Spinal Disease Research, Beijing 100191, China
[5]Department of Urology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
[6]Department of Head and Neck Surgery, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
[7]Department of Thoracic Surgery II, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
[8]Sun Yat-sen University Cancer Centre, State Key Laboratory of Oncology in South China, Collaborative Innovation Centre for Cancer Medicine, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangzhou 510060, China
[9]Department of Gastrointestinal Surgery, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
[10]Department of Biobank, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
[11]Institute of Cancer Research, Shenzhen Bay Laboratory, Shenzhen 518132, China
[12]Peking University International Cancer Institute, Beijing 100191, China
[13]These authors contributed equally
[14]Lead contact
*Correspondence: buzhaode@cjcrcn.org (Z.B.), jijiafu@hsc.pku.edu.cn (J.J.), zemin@pku.edu.cn (Z.Z.)
https://doi.org/10.1016/j.cell.2021.01.010

## SUMMARY

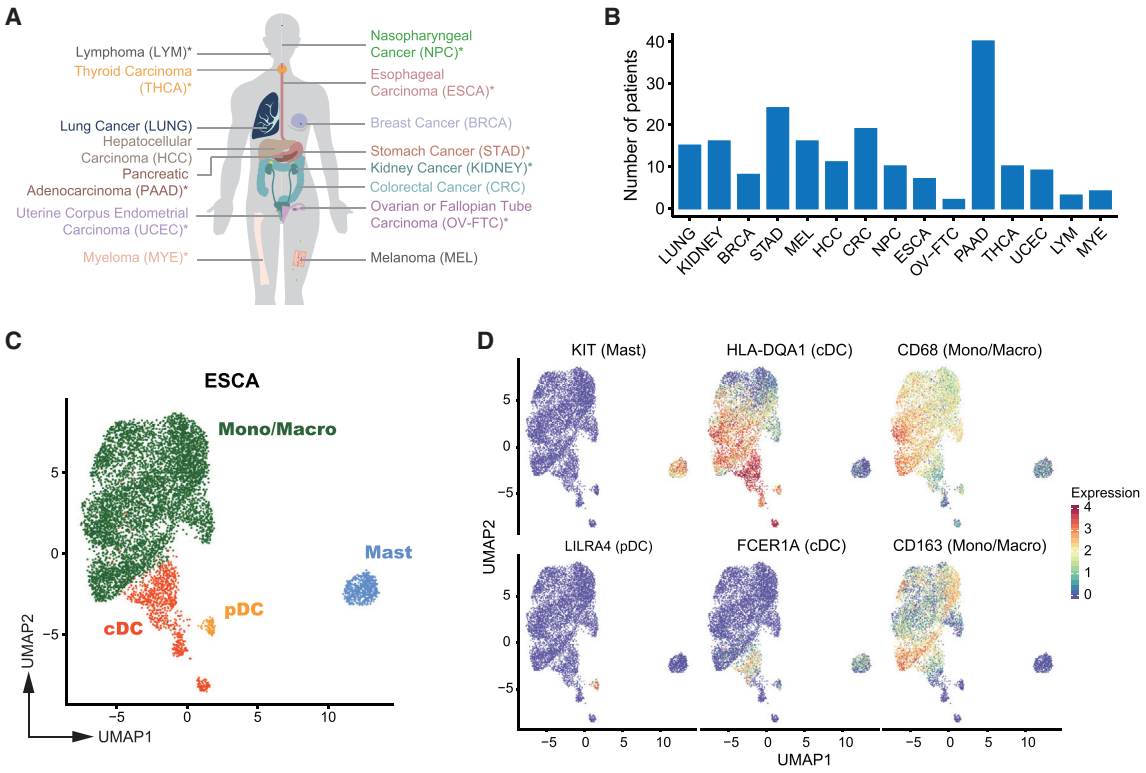
Tumor-infiltrating myeloid cells (TIMs) are key regulators in tumor progression, but the similarity and distinction of their fundamental properties across different tumors remain elusive. Here, by performing a pan-cancer analysis of single myeloid cells from 210 patients across 15 human cancer types, we identified distinct features of TIMs across cancer types. Mast cells in nasopharyngeal cancer were found to be associated with better prognosis and exhibited an anti-tumor phenotype with a high ratio of *TNF*$^{+}$/*VEGFA*$^{+}$ cells. Systematic comparison between cDC1- and cDC2-derived *LAMP3*$^{+}$ cDCs revealed their differences in transcription factors and external stimulus. Additionally, pro-angiogenic tumor-associated macrophages (TAMs) were characterized with diverse markers across different cancer types, and the composition of TIMs appeared to be associated with certain features of somatic mutations and gene expressions. Our results provide a systematic view of the highly heterogeneous TIMs and suggest future avenues for rational, targeted immunotherapies.

## INTRODUCTION

Tumors are complex ecosystems where heterogeneous malignant cells interact with both immune and nonimmune cells to shape the complex cellular network of the tumor microenvironment (TME) (Hanahan and Weinberg, 2011). Because myeloid cells constitute a key cellular component of immune cells that infiltrate into tumors and play important roles in modulating tumor inflammation and angiogenesis (Engblom et al., 2016; Schmid and Varner, 2010), several therapeutic approaches targeting myeloid cells are ongoing in pre-clinical and clinical studies (Nakamura and Smyth, 2020), although the heterogeneity of myeloid cells remains less studied.

TIMs consist of several distinct major lineages including mast cells, plasmacytoid dendritic cells (pDCs), conventional dendritic cells (cDCs), monocytes, and macrophages (Engblom et al., 2016). In the past decade, primarily with the aid of flow cytometry, the complexity of major myeloid lineages has begun to be revealed. Monocytes are usually classified based on the expression of surface markers CD14 and CD16 (Yang et al., 2014).




*(legend on next page)*

Macrophages are critical mediators in TME and participate in multiple aspects of tumor immunity (DeNardo and Ruffell, 2019). The "classically activated" M1 and "alternatively activated" M2 macrophage polarization system has been used to describe the *in vitro* activation state of macrophages (Vogel et al., 2014). However, macrophages *in vivo* exhibit more complex phenotypes, which argue against such a simple categorization *in vitro* (Ginhoux et al., 2016). Dendritic cells (DCs) are key players in antigen-specific immune responses (Mellman, 2013). Two distinct cDC subsets, XCR1$^+$CADM1$^+$ cDC1s and CD1A$^+$ CD172A$^+$ cDC2s, have been identified and shown to interact with CD8$^+$ and CD4$^+$ T cells, respectively (Binnewies et al., 2019; Salmon et al., 2016). Furthermore, the heterogeneity of cDC2s only starts to be revealed recently, as distinct cDC2 subsets in human blood and spleen have been identified (Brown et al., 2019; Dutertre et al., 2019; Villani et al., 2017). However, the complexity of cDC2s across tumors is still not fully characterized.

The single-cell RNA sequencing (scRNA-seq) technologies have been increasingly applied to characterize TME at the single-cell resolution (Ren et al., 2018). At present, several studies have delineated the diversity of TIMs in a single cancer type when characterizing the global immune landscape of TME (Azizi et al., 2018; Sharma et al., 2020; Zhang et al., 2019, 2020; Zilionis et al., 2019), although some of them did not fully characterize the complexity of TIMs. Multiple interesting myeloid cell subsets have been identified, including the *LAMP3*$^+$ cDCs—a relatively under-characterized mature cDC subset highlighted in hepatocellular carcinoma (HCC) (Zhang et al., 2019), the *LYVE1*$^+$ resident tissue macrophages (RTMs) related to restraining inflammation and fibrosis in multiple human tissues (Chakarov et al., 2019) and colon cancer (CRC) (Zhang et al., 2020), the *SPP1*$^+$ TAMs associated with tumor angiogenesis reported in CRC (Zhang et al., 2020), and the *FOLR2*$^+$ TAMs supporting the onco-fetal reprogramming of TME in HCC (Sharma et al., 2020). However, it is not clear whether such findings can be extended to other cancer types, and it remains unknown whether the same subsets of TIMs are present across different cancer types.

In this study, using a combination of published and newly generated scRNA-seq data, we painted the landscapes of TIM across 15 tumor types, characterized the degree of similarity of myeloid cell subsets across cancer types, and investigated the unique composition and characteristics of those subsets in different cancer types. Our comprehensive TIM compendium would provide an unprecedented resource to understand the complexity of the TME and guide immunotherapy development for multiple human cancer types.

## RESULTS

### Landscapes of myeloid cells in 15 cancer types revealed by scRNA-seq analysis

To generate a deep transcriptional atlas of TIMs, we obtained scRNA-seq data on myeloid cells in 380 samples from 210 patients diagnosed with one of the 15 common cancer types, including newly collected 82 treatment-naive patients of 10 cancer types (Figure 1A; Table S1). After strict quality control and filtration, we collected a total of 138,161 myeloid cells derived from the tumors, adjacent non-cancer tissues, peripheral blood, or lymph node of 194 patients (338 samples) across 15 common cancer types (Figures 1A, 1B, S1A, and S1B; Table S1; STAR methods).

To characterize the subsets of myeloid cells and minimize batch effects among different datasets, we analyzed each dataset independently. We performed unsupervised graph-based clustering on myeloid cells and then identified four common major linages (mast cells, pDCs, cDCs, monocytes, or macrophages) based on canonical cell markers. In addition, cDCs and monocytes or macrophages could be further divided into multiple sub-populations (Figure S1C, STAR methods). Using esophageal carcinoma (ESCA) as an example, mast cells, pDCs, cDCs, and monocytes/macrophages were characterized by specific high expression of *KIT*, *LILRA4*, *HLA/FCER1A*, and *CD68/CD163*, respectively (Figures 1C and 1D). Three distinct subsets in cDCs were identified (Figures 1E and S1C), including two classical cDC subsets (*CLEC9A*$^+$ cDC1s and *CD1C*$^+$ cDC2s) and a mature cDC subset (*LAMP3*$^+$ cDC) recently characterized (Zhang et al., 2019). Further clustering of the monocytes/macrophages in ESCA gave rise to 6 sub-populations with specific gene signatures, including two groups of monocytes (classical CD14$^{hi}$CD16$^-$ and non-classical CD14$^+$CD16$^{hi}$) and four groups of macrophages (Figures 1E and S1C). Likewise, cDCs and monocytes in all other 14 cancer types were clustered into the same sub-populations with consistent markers (Figures 1F and S1C). Of note, in CRC and myeloma, we identified an intermediate monocyte subset (CD14$^{hi}$CD16$^+$ monocytes) (Yang et al., 2014), which was likely to benefit from the large number of monocytes in the blood. In contrast to cDCs and monocytes, macrophage subsets in different tumor types exhibited much higher heterogeneity (Figures 1F and S1C). For example, we identified a lung cancer-specific macrophage population Macro_PPARG, which represented lung-resident alveolar macrophages, characterized by high expression of *PPARG*, *MARCO*, *MRC1*, and *MSR1* (Gordon et al., 2014; Figures S1C and S2A). The rationality of our cluster identification was further verified using the four

**Figure 1. Identification of myeloid cell subsets in various cancer types**

(A) The cancer types included in the pan-cancer analysis. *, cancer types with newly generated data.

(B) The number of patients collected for each cancer type.

(C) UMAP plot showing the major lineages of myeloid cells in ESCA.

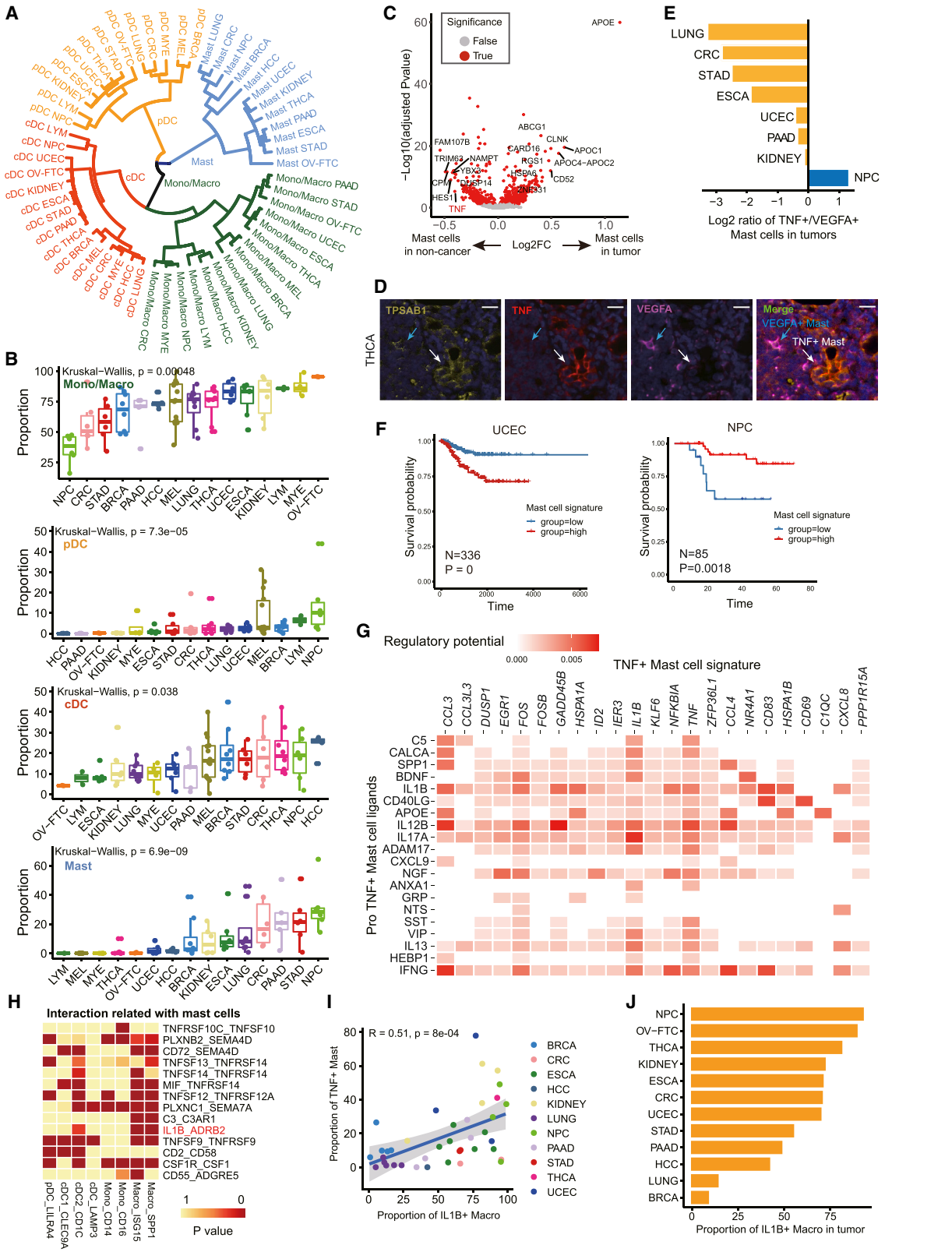
(D) UMAP plots showing the marker genes expression for the major lineages of myeloid cells.

(E) Bubble heatmap showing expression levels of selected signature genes in ESCA. Dot size indicates fraction of expressing cells, colored based on normalized expression levels.

(F) Heatmap showing the presence of different myeloid cell subsets.

(G) UMAP plots showing the integration of two LUNG scRNA-seq datasets.

See also Figures S1 and S2, and Table S1.




*(legend on next page)*

cancer types comprising two independent studies (Figures 1G and S2B-S2D, STAR methods).

## Mast cells exhibited diverse functional potentiality in different cancer types

We next compared the relationship between the major lineages of myeloid cells across different cancer types. We first integrated the large number of myeloid cells across 15 cancer types and observed that cells with the same major lineage identities were well mixed (Figure S3A). To quantify their similarities, we calculated the correlations between the average transcriptome of each major lineage in different cancer types. As expected, the same major lineages from different cancer types were clustered together (Figures 2A and S3B, STAR methods), further demonstrating that major myeloid lineages shared similar transcriptomic profiles.

We then examined the composition of the major lineages in each tumor type (Figure 2B). Monocytes and macrophages accounted for the largest proportion of TIMs, with an average of above 50% in most tumors. The pDCs lineage was a small sub-population at tumor site, comprising less than 10% of TIMs. The proportion of cDCs was relatively stable (~10%–20%) across tumor types. By contrast, the proportion of mast cells showed dramatic variation across different tumors. For instance, mast cells were largely absent in uterine corpus endometrial carcinoma (UCEC) and HCC, whereas the ratio of mast cells was relatively high in nasopharyngeal cancer (NPC), which hinted that mast cells might exhibit diverse functions across cancer types.

We further examined the tissue distribution of mast cells. Higher proportions of mast cells in tumor tissues were observed in most cancer types (Figure S3C), supporting that mast cells accumulate in tumors and play important roles in tumorigenesis and tumor progression (Maciel et al., 2015). However, we noticed that mast cells increased in the adjacent normal uterus (Figure S3D), which was further confirmed by the *Ro/e* analysis (Figure S3E; STAR methods) and could be associated with reproductive processes (Woidacki et al., 2013). Interestingly, lower expression of *TNF* was observed in mast cells from tumor tissues compared with that from non-cancer tissues in UCEC (Figure 2C). Mast-derived *TNF* has been reported to play a critical role in DC functionality and T cell priming, which is beneficial for anti-tumor capacity (Dudeck et al., 2015). We then extended our analysis to compare the expression patterns of *TNF* and found that tumor-derived mast cells showed low expression of *TNF* in multiple cancer types (Figure S3F), suggesting the presence of an inhibitory program that hinders the anti-tumor activities of mast cells in those tumors. Moreover, we observed an increase of $TNF^+$ mast cells in lung cancer (LUNG) patients at the residual disease (RD) state, compared to patients before tyrosine kinase inhibitors therapy (TN) or at the progressive disease (PD) state (Figure S3G; Maynard et al., 2020), which further supports the importance of $TNF^+$ mast cells in anti-tumor immunity.

Mast cells in tumors are thought to play a dual role in influencing the fate of tumor cells (Ribatti, 2016). Apart from *TNF*, mast-derived *VEGFA* is a key gene associated with tumor angiogenesis (Marone et al., 2016), which informed us to evaluate the complexity of mast cells. Using ROGUE analysis (Liu et al., 2020), we observed that mast cells showed the highest homogeneity among the four major lineages (Figure S3H). We then directly compared the expression patterns of *TNF* and *VEGFA* in mast cells and observed both mutually exclusive and co-expression pattern across different cancer types (Figure S3I), suggesting that masts cell in different tumors might exhibit different compounding effects on tumor cells. Immunohistochemistry (IHC) staining of tumor sections further confirmed the existence of these two subsets of mast cells in thyroid carcinoma (THCA) (Figure 2D). To compare their effects on tumor cells across cancer types, we quantified the ratio of $TNF^+$ to $VEGFA^+$ mast cells in tumors and observed that the frequencies of $VEGFA^+$ mast cells were much higher than $TNF^+$ cells in most tumor types (Figure 2E), indicating their prevailing pro-angiogenic role. Intriguingly, NPC stood up as the only cancer type with dramatically higher frequency of $TNF^+$ than $VEGFA^+$ mast cells (ratio = 2.4) (Figure 2E), indicating a higher anti-tumor capacity. Using clinical data collected from the TCGA project (Liu et al., 2018) and NPC patients, we confirmed that the higher expression of mast cell markers was associated with a survival advantage in NPC, while an opposite effect was observed in multiple cancer types (Figures 2F and S3J; STAR methods), consistent with our hypothesis.

**Figure 2. Characterization of major myeloid cell lineages in various cancer types**

(A) Hierarchical clustering of major myeloid cell lineages.

(B) Boxplots showing the proportions of each major lineage divided by the total myeloid cell number in tumor tissues across various cancer types. Kruskal-Wallis test.

(C) Volcano plot showing differentially expressed genes between mast cells from tumor tissues and non-cancer tissues in UCEC. Adjusted p value <0.05, Benjamini-Hochberg adjusted two-sided Wilcoxon test.

(D) Representative example of a thyroid tumor stained by IHC with anti-TPSAB1 (yellow), TNF (red), and VEGFA (violet) antibodies. Arrows depict the specific cell types. Scale bar, 20 μm.

(E) Bar plot showing the ratios of $TNF^+$ mast cells to $VEGFA^+$ mast cells in multiple cancer types (with more than 200 mast cells).

(F) Kaplan-Meier plots showing worse clinical outcome in UCEC and better clinical outcome in NPC patients with the higher expression of mast cell marker genes. +, censored observations; HR, hazard ratio. Cox regression.

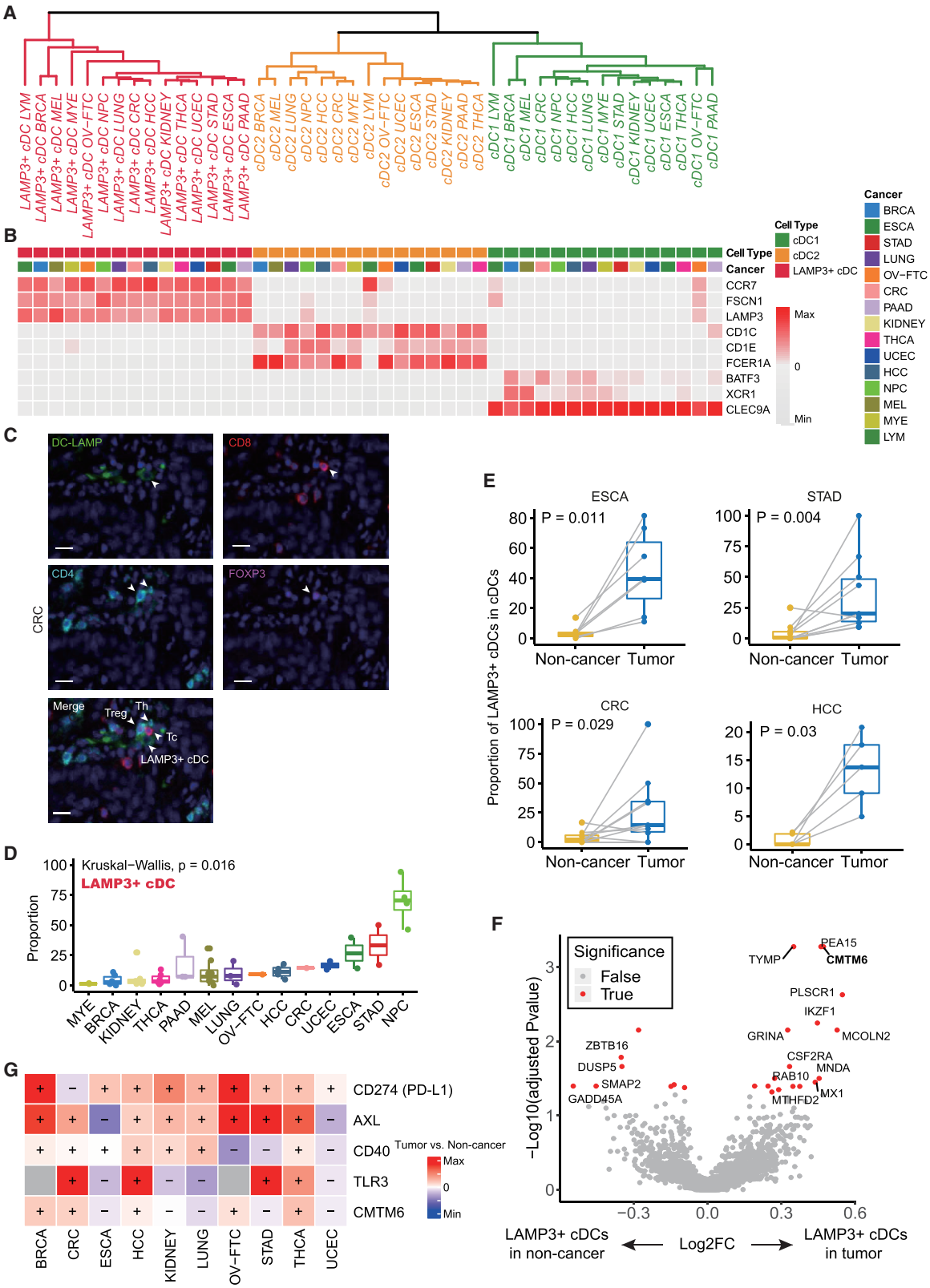
(G) Heatmap showing potential ligands driving the phenotype of $TNF^+$ mast cells.

(H) Heatmap showing the selected ligand-receptor pairs between mast cells and other myeloid cells in NPC.

(I) Scatterplot showing the Pearson correlation of the proportion of $IL1B^+$ macrophages (divided by the total macrophage number) and $TNF^+$ mast cells (divided by the total mast cell number) in tumor tissues.

(J) Bar plot showing the proportion of $IL1B^+$ macrophages divided by the total macrophage number.

See also Figure S3.




*(legend on next page)*

Taken together, these findings reveal unique anti-tumor function of mast cells in NPC, whereas mast cells usually play a pro-tumorigenic role in most other tumors.

To explore the underlying reasons for the enrichment of *TNF*$^+$ mast cells in NPC, we compared the transcriptomic differences between *TNF*$^+$ and *TNF*$^-$ mast cells, and then used the highly expressed genes in *TNF*$^+$ mast cells for NicheNet analysis (Browaeys et al., 2020; STAR methods). We noticed that SPP1 and interleukin (IL)-1β, highly produced by TAMs, were predicted as the possible ligands to drive the phenotype of *TNF*$^+$ mast cells (Figure 2G). Using CellphoneDB (Efremova et al., 2020), we also found that mast cells were more likely to interact with macrophages in NPC (Figure S3K), and one of the significantly enriched ligand-receptor pairs was IL1B-ADRB2 (Figure 2H). ADRB2 has been reported to possess a critical role in preventing the release of pro-inflammatory mediators from mast cells (Barnes, 1999). Furthermore, we identified a positive correlation between *IL1B*$^+$ macrophages and *TNF*$^+$ mast cells in tumor tissues (Figure 2I). Meanwhile, NPC was found to harbor the highest proportion of *IL1B*$^+$ macrophages (Figure 2J). Taken together, these results indicate that the presence of *IL1B*$^+$ macrophages may drive the anti-tumor property of mast cells in NPC.

## Distinct developmental origins of the widely present *LAMP3*$^+$ cDCs

Next, we focused on the heterogeneity and functions of cDCs. To assess the similarity of the three cDC subsets across different tumor types, we observed that the same cDC subsets from different cancer types were clustered together (Figures 3A and 3B), suggesting that the transcriptomic states of each cDC subset are developmentally determined, with cancer types having relatively marginal impacts. We previously reported that the *LAMP3*$^+$ cDCs represented a distinct subset of cDCs in HCC (Zhang et al., 2019). Here, we identified the *LAMP3*$^+$ cDCs in all 15 cancer types (Figures 1F and 3A), indicating their broad presence in TME. IHC staining of tumor sections from CRC patients further confirmed the existence of *LAMP3*$^+$ cDCs and showed the physical juxtaposition of *LAMP3*$^+$ cDCs and multiple T types (Figure 3C). In line with the characteristics reported in HCC (Zhang et al., 2019), the *LAMP3*$^+$ cDCs in all cancer types exhibited an "activated" phenotype with higher capacity of migratory ability (Figure S4A; Table S2). Deep inspection of the signature genes and the similarity analysis indicated that the *LAMP3*$^+$ cDCs closely resembled those "mregDCs" (Maier et al., 2020; STAR methods), a cluster of DCs showing high expression of maturation markers, regulatory molecules, and migratory genes, as well as low expression of Toll-like receptor (TLR) signaling genes (Figures S4B and S4C). Altogether, these findings indicate that the *LAMP3*$^+$ cDCs represent mature cDCs and are widely present in diverse tumors, implying their indispensable roles in the TME.

We next examined the composition of cDC subsets across cancer types. Overall, we observed a higher proportion of cDC2s than cDC1s in tumors (Figure S4D) and the abundance of *LAMP3*$^+$ cDCs varied widely in different cancer types (Figure 3D). Compared with adjacent non-cancer tissues, a much higher proportion of the *LAMP3*$^+$ cDCs in tumor tissues was observed in multiple types of cancer (Figures 3E, S4E, and S4F). However, in THCA, we noticed that the *LAMP3*$^+$ cDCs tended to be enriched in adjacent non-cancer tissues (Figure S4G), which inspired us to examine the differentially expressed genes between *LAMP3*$^+$ cDCs from tumor with those from non-cancer tissues. One of the highly expressed genes in tumor-derived *LAMP3*$^+$ cDCs was *CMTM6* (Figure 3F), a key PD-L1 protein regulator by reducing its ubiquitination and increasing PD-L1 protein half-life (Burr et al., 2017; Mezzadra et al., 2017). As reported, *LAMP3*$^+$ cDCs could potentially interact with multiple T cell subsets via PD-1/PD-L1 (Zhang et al., 2019), leading us to compare the expression of *PD-L1* and its regulators in *LAMP3*$^+$ cDCs. We found that *PD-L1* was universally upregulated in tumor-derived *LAMP3*$^+$ cDCs in almost all cancer types, whereas its regulators, including *AXL*, *TRL3*, and *CD40* (Maier et al., 2020) showed distinct patterns (Figure 3G). Taken together, our data suggest that although the regulation of *PD-L1* in tumor-derived *LAMP3*$^+$ cDCs is complex and mediated by diverse regulators, the upregulation of *PD-L1* is consistent. It remains to be determined whether the efficacy of anti-PD-L1 antibodies is associated with the expression level of *PD-L1* in the tumor-derived *LAMP3*$^+$ cDCs.

Recent studies have suggested that *LAMP3*$^+$ cDCs had the potential to develop from both cDC1s and cDC2s (Maier et al., 2020; Zhang et al., 2019). Using multiple computational methods, including Monocle2 (Qiu et al., 2017), CytoTRACE (Gulati et al., 2020), the direct stratification of *LAMP3*$^+$ cDCs and SingleR (Aran et al., 2019; STAR methods), we confirmed the two potential origins of the *LAMP3*$^+$ cDCs among various cancer types (Figures 4A, 4B, and S5A–S5C). Notably, although the proportion of cDC2s was much higher in tumor tissues, we observed more cDC1-derived *LAMP3*$^+$ cDCs in all cancer types, except for pancreatic adenocarcinoma (PAAD) and NPC (Figures 4C and S5D), which was further confirmed by an independent PAAD dataset (Peng et al., 2019; Figure S5E). We reasoned

**Figure 3. Characterization of cDC subsets in various cancer types**

(A) Hierarchical clustering of cDC subsets.

(B) Heatmap showing the common markers for distinct cDC subsets.

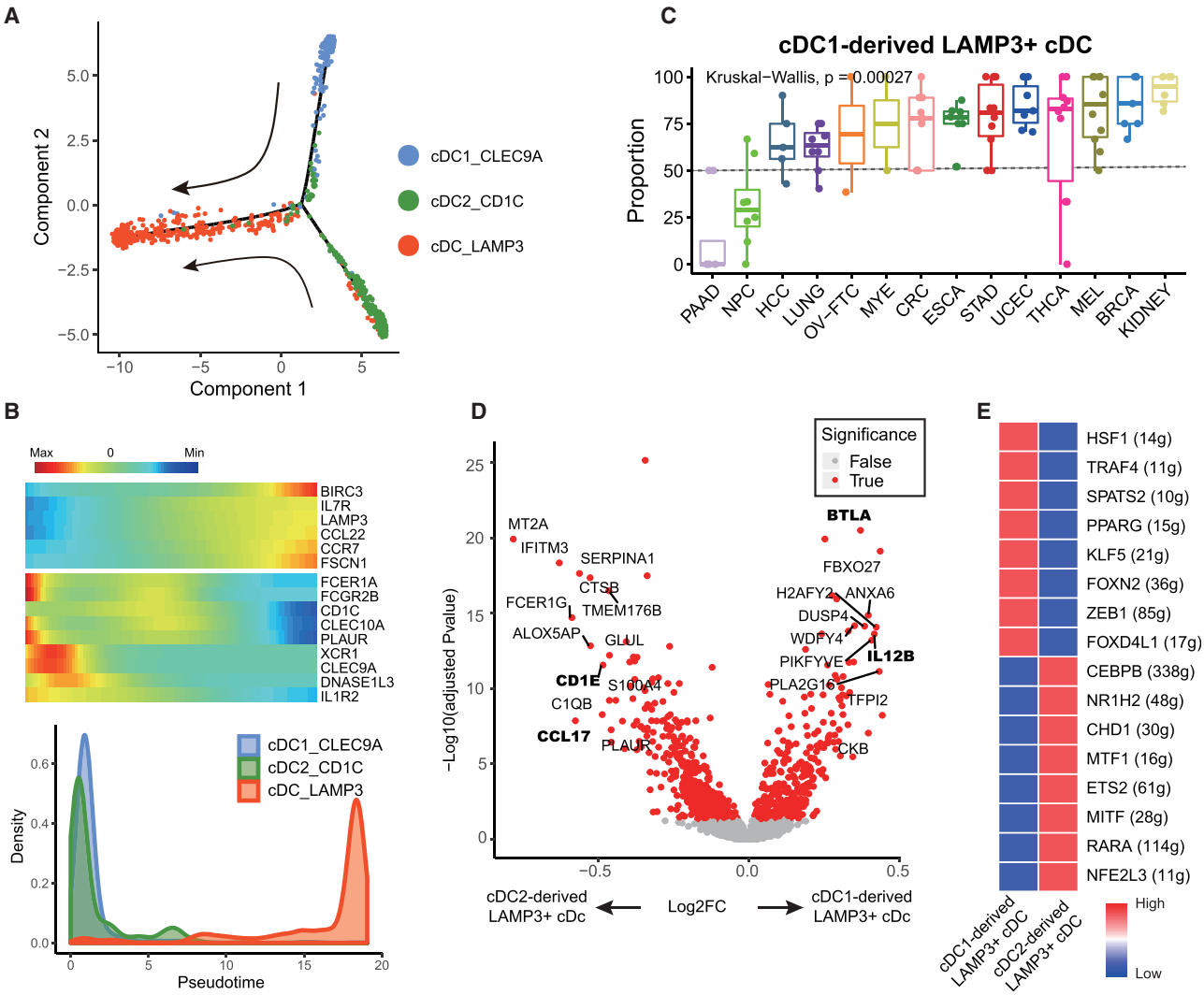
(C) Representative example of a CRC tumor stained by IHC with anti-CD8 (red), CD4 (cyan), FOXP3 (violet), and DC-LAMP (green) antibodies. Arrows depict the specific cell types. Scale bar, 20 μm. Th, T helper; Tc, cytotoxic T cell; Treg, regulatory T cell.

(D) Boxplots showing the proportions of *LAMP3*$^+$ cDC subset (divided by the total cDC number) in tumor tissues across various cancer types (with more than 5 *LAMP3*$^+$ cDCs). Kruskal-Wallis test.

(E) Boxplots showing the fraction of *LAMP3*$^+$ cDCs (divided by the total cDC number) in tumor and non-cancer tissues. Paired two-sided Wilcoxon test.

(F) Volcano plot showing differentially expressed genes between *LAMP3*$^+$ cDCs from tumor tissues and non-cancer tissues in THCA. Adjusted p value <0.05, Benjamini-Hochberg adjusted two-sided Wilcoxon test.

(G) Heatmap showing the fold change of gene expression between tumor-derived and non-cancer-derived *LAMP3*$^+$ cDCs. Missing value was colored in gray. See also Figure S4 and Table S2.




(legend on next page)

that the higher proportion of cDC2-derived *LAMP3*$^+$ cDCs in PAAD might interact with tumor-infiltrating Treg cells or inhibit CD8$^+$ T cell-mediated tumor immunity to promote immune tolerance and metastasis (Jang et al., 2017; Kenkel et al., 2017).

To study whether *LAMP3*$^+$ cDCs from different origins might have different functions, we examined their transcriptomic differences (Table S3). Consistent with previous finding (Maier et al., 2020), *IL12B*, which could induce the differentiation of T helper 1 ($T_H1$) cells (Trinchieri, 2003), was specific to cDC1-derived *LAMP3*$^+$ cDCs (Figures 4D and S5F). cDC1-derived *LAMP3*$^+$ cDCs also showed higher expression of *BTLA* (Figures 4D and S5F), which could induce Treg differentiation and result in immune tolerance (Simon and Bromberg, 2016), in line with the capacity of mregDCs to drive the differentiation of naive T cells into Treg cells (Maier et al., 2020). By contrast, cDC2-derived *LAMP3*$^+$ cDCs still maintained high expression of cDC2 marker gene *CD1E* and showed higher expression of *CCL17* (Figures 4D and S5F), a chemokine that could recruit *CCR4*$^+$ Treg into tumor and create immunosuppressive environment (Ishida and Ueda, 2006). Moreover, using SCENIC (Aibar et al., 2017) to detect regulatory networks, we uncovered a series of regulons that underlie each *LAMP3*$^+$ cDCs (Figure 4E), including known DC-related transcription factor RARA (Hashimoto-Hill et al., 2018) and ZEB1 (Scott and Omilusik, 2019). These findings imply that, although the two cDC subsets could converge after differentiating into *LAMP3*$^+$ cDCs, they maintain specific transcriptomic properties, which might diversify their functions.

We then explored extracellular signals that drive the maturation of cDCs and cataloged three distinct gene signatures: cDC1-specific, cDC2-specific and shared maturation signatures (Table S4; STAR methods). Ligands identified from the shared maturation signatures (Figure S5G) may activate the conserved maturation and activation processes shared by cDC1s and cDC2s. For instance, IL-21 has been reported to modify DC ability and then selectively enhance NKT cell production of interferon (IFN)-$\gamma$ (Maeda et al., 2007). Additionally, we also identified specific ligands for each cDC subset (Figure 4F). Of the ligands regulating the maturation of cDC1s, IL-4 and *IL12B* were significantly enriched, consistent with previous studies (Maier et al., 2020), whereas IL-15, a cytokine reported to induce the conversion of monocytes to mature DCs (Saikh et al., 2001), was identified from cDC2-specific maturation signatures. Interestingly, multiple IFNs were prioritized in cDC2s, suggesting that they might play important roles in the maturation of cDC2s. Collectively, our data suggest that the maturation of DCs are simultaneously regulated by a set of conserved ligands and cDC1/cDC2-specific ligands. Further studies will be needed to validate the functional roles of these putative regulators and to fully understand the diverse differentiation processes of cDC subsets.

## Variation of cDC2 sub-populations across tumors

We next used our newly generated scRNA-seq data from 8 cancer types to interrogate the heterogeneity of cDC2s. We first applied ROGUE and found that cDC2s exhibited the highest heterogeneity among four DC populations (Figure 5A). Unsupervised graph-based clustering gave rise to 6 cDC2 clusters with specific signature genes (Figures 5B–5D and S6A; Table S4). Among these clusters, cells of C01_cDC2_CXCR4 cluster were almost evenly distributed across different tissues. The C02_cDC2_CD1A cluster, predominantly composed of cells from non-blood tissues, specifically expressed Langerhans cell-specific markers: *CD1A* and *CD207* (langerin), but not *EPCAM*, suggesting that this cluster represented a subset of langerin$^+$ DCs (Merad et al., 2008). The tissue-enriched C03_cDC2_IL1B and blood-enriched C04_cDC2_FCN1 clusters both showed lower expression of *CD5* and intermediate or high expression of both *CD163* and *CD14*, resembling the reported "pro-inflammatory" cDC2 sub-population in blood (Dutertre et al., 2019). Notably, the C03_cDC2_IL1B cluster showed high expression of pro-inflammatory mediators (*IL1B* and *TNF*) and T cell chemokines (*CCL3* and *CCL4*), which further supported their "pro-inflammatory" properties and functions in CD8$^+$ T cells recruitment (Castellino et al., 2006). The C05_cDC2_ISG15 cluster had high expression levels of multiple interferon inducible genes, such as *ISG15*, *IFI6*, and *IFIT1*. The remaining cDC2 cells, falling into the C06_cDC2_CXCL9 cluster, were dominant in tumor tissues and were characterized by specific expression of chemokine *CXCL9*.

We then examined the regulatory network that underlies each cDC2 subset using SCENIC, and identified specific TF regulons for each cDC2 subset (Figure 5E). For example, the langerin$^+$ C02_cDC2_CD1A subset tuned up the positive regulators SPI1 (PU.1) and RUNX3 and downregulated the negative C/EBP regulons (Figures 5E and S6B), which have been reported to play critical roles in the development and maintenance of Langerhans cell (Zhang et al., 2016). Specifically, the PRDM1 and BCL6 regulons, known to be critical regulators of effector and memory differentiation of T and B lymphocytes (Crotty et al., 2010), were highly activated in the C03_cDC2_IL1B cluster, suggesting its potential to promote T cell differentiation. Taken together, these analyses identified plausible candidates to drive the phenotypic

**Figure 4. Comparison of cDC1- and cDC2-derived *LAMP3*$^+$ cDCs**

(A) Developmental trajectory of cDC subsets inferred by Monocle2.

(B) The patterns of gene expression (top) and cell density (bottom) along with the pseudotime.

(C) Boxplot showing the fractions of cDC1-derived *LAMP3*$^+$ cDCs (divided by the total *LAMP3*$^+$ cDC number) across various cancer types (with more than 5 *LAMP3*$^+$ cDCs). Kruskal-Wallis test.

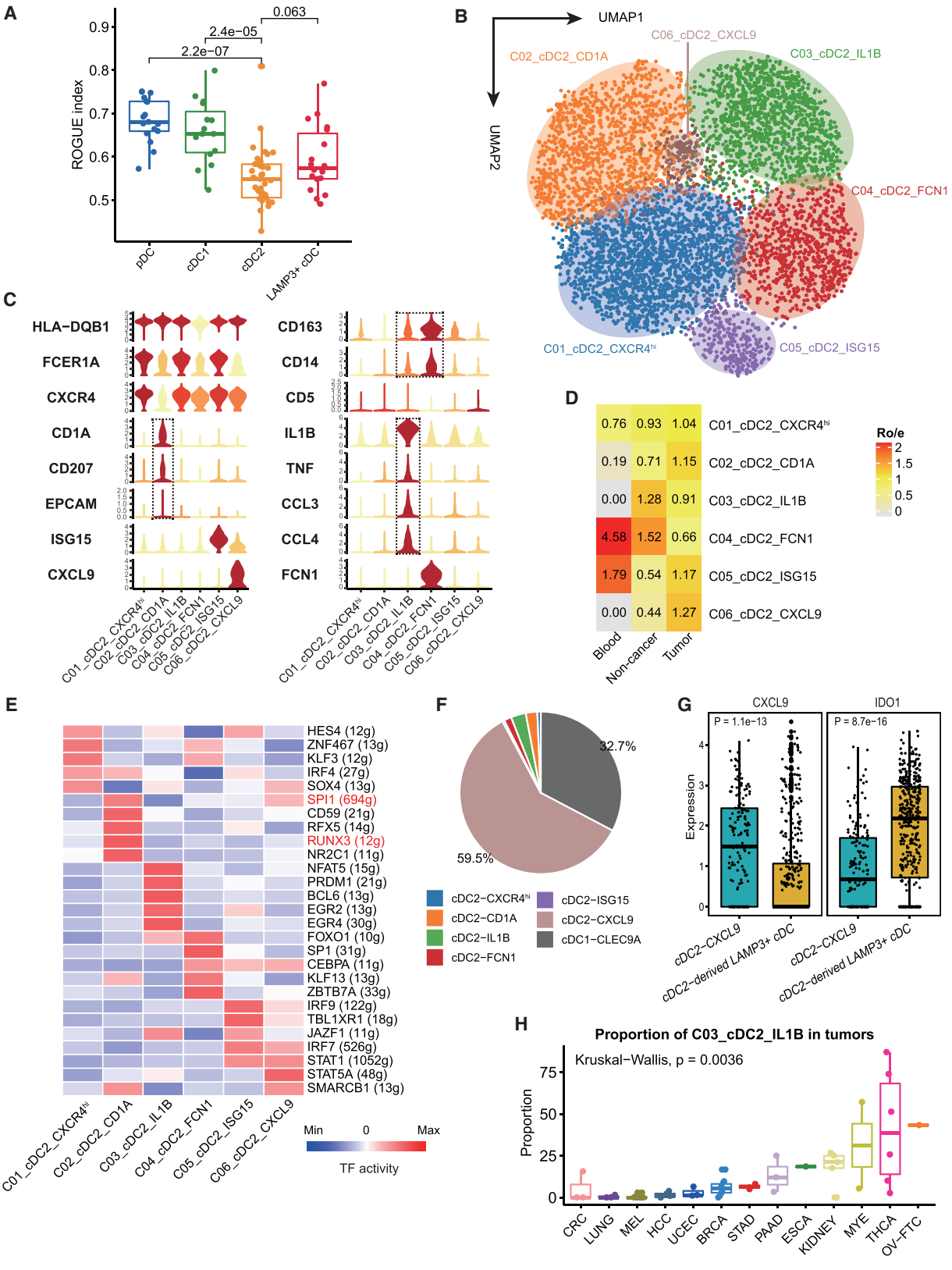
(D) Volcano plot showing differentially expressed genes between cDC1-derived and cDC2- derived *LAMP3*$^+$ cDCs. Adjusted p value <0.05, Benjamini-Hochberg adjusted two-sided Wilcoxon test.

(E) Heatmap showing TF activity for cDC1- and cDC2-derived *LAMP3*$^+$ cDCs.

(F) Heatmaps showing potential ligands driving the maturation of cDC1s (left) and cDC2s (right).

(A), (B), and (D)–(F) were plotted using the integrated data from newly generated datasets in our lab.

See also Figure S5 and Tables S3 and S4.




*(legend on next page)*

differences of each cDC2 subset and could help elucidate the heterogeneity and function of these cells.

We further explored the relationship between the cDC2 subsets and *LAMP3*$^+$ cDCs to determine which cDC2 subset was more likely to differentiate into *LAMP3*$^+$ cDCs. Among these cDC2 subsets, the C06_cDC2_CXCL9 sub-cluster was predicted as one of the main origins of *LAMP3*$^+$ cDCs (Figures 5F and S6C). Comparing with other cDC2 subsets, the C06_cDC2_CXCL9 sub-cluster showed higher expression of *CXCL9* and *IDO1* (Figures S6D and S6E), which have been reported to regulate immune activation and induce immune repression, respectively (Tokunaga et al., 2018; Wu et al., 2018). Interestingly, the expression of *CXCL9* was downregulated and the expression of *IDO1* was upregulated during the transition from *CXCL9*$^+$ cDC2s to *LAMP3*$^+$ cDC (Figure 5G), suggesting that the *CXCL9*$^+$ cDC2s acquire an enhanced immunosuppressive functionality when developing into cDC2-derived *LAMP3*$^+$ cDCs.

We next compared the cDC2 subset composition in tumors (Figure S6F; STAR methods) and noticed that the fraction of the two pro-inflammatory (*FCN1*$^+$ and *IL1B*$^+$) cDC2 subsets exhibited significant variations across different tumor types (Figures 5H and S6G). The results indicated that those cDC2 subsets might contribute differently to tumor immune responses, and different responses would be observed across various cancer types when applying therapeutic strategies targeting those pro-inflammatory cDC2 subsets.

## Macrophage subsets exhibit distinct transcriptomic patterns among various tumor types

To compare the relationships between sub-populations of monocytes and macrophages, we performed the similarity analysis as outlined above. Both CD14$^+$ and CD16$^+$ monocytes from different cancer types were clustered into two branches mainly separated by tissue sources (Figures 6A and S7A), indicating that monocyte subsets were less affected by local tissue ecosystem. We then investigated whether these monocytes were re-programed by TME. Although all monocytes showed high expression of classical monocyte related genes (*CD14*, *CD16*, and *FCN1*), blood-derived monocytes showed higher expression of a set of neutrophil-associated genes (*S100A8*, *S100A9*, and *CSF3R*), whereas tumor or adjacent normal tissue-derived monocytes showed higher expression of human leukocyte antigen (HLA) genes (*HLA-DQA1* and *HLA-DQB1*) and macrophage related genes (*CD163* and *CD68*) (Figure S7B), suggesting an intermediate state of monocytes migrating into tissues and maturing into macrophages. Furthermore, tumor-infiltrating monocytes expressed higher levels of inflammatory cytokines and chemokines (*IL1B*, *CCL4*, *CXCL2*, and *CXCR4*), cell growth regulators (*AREG* and *EGR1*), tissue resident markers (*NR4A1*, *NR4A2*, *NLRP3*), and nuclear factor κB (NF-κB) signaling genes (*NFKB1* and *NFKBIA*) (Figure S7B).

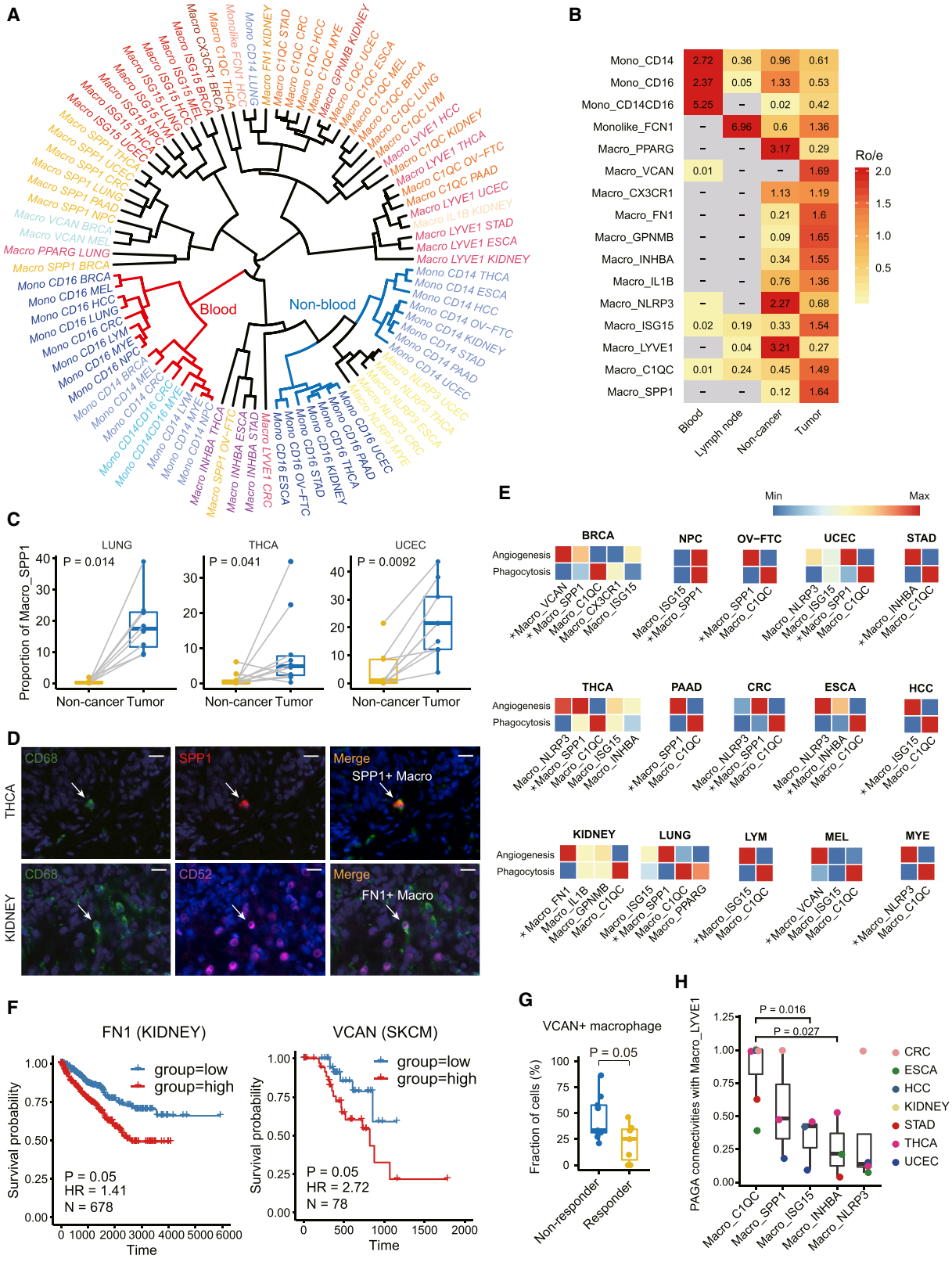
By contrast, diverse macrophage subsets across different cancer types were identified (Figures 1F and S1C). Notably, the similarity analysis failed to exactly cluster macrophage subsets with the same identity from different cancer types together (Figure 6A), indicating that macrophage subsets exhibited high level of complexity, which might be related to the dominant effects of the local tissue microenvironment on macrophages (Gosselin et al., 2014; Lavin et al., 2014). These findings were further confirmed by an entropy-based metric to quantify the mixability of the data and Scanorama integration (Figures S7C and S7D; STAR methods). Likewise, based on close inspection of the expression of marker genes *SPP1* and *C1QC* highlighted by our recent publication (Zhang et al., 2020), we found that *SPP1* tended to have a mutually exclusive relationship with *C1QC* in CRC and breast cancer (BRCA), but co-expressed with *C1QC* in a subset from UCEC (Figure S7E). Thus, macrophage subsets from different cancer types showed heterogeneous transcriptomic patterns and need to be analyzed separately to better describe their properties.

We next quantified tissue enrichment of macrophage subsets based on the *Ro/e* analysis (Figures 6B and S7F). The *LYVE1*$^+$ macrophages identified within multiple cancer types were preferentially enriched in non-cancer tissues, resembling the reported tissue-resident interstitial macrophages (Chakarov et al., 2019). Likewise, the *NLRP3*$^+$ macrophages were also enriched in non-cancer tissues, likely representing a cluster of pro-inflammatory TRMs. Furthermore, we identified several tumor-enriched macrophage subsets, which were denoted as tumor-associated macrophages (TAMs), including the *SPP1*$^+$ TAMs (Figure 6C), *C1QC*$^+$ TAMs, *ISG15*$^+$ TAMs, and *FN1*$^+$ TAMs. The *ISG15*$^+$ TAMs upregulated multiple interferon-inducible genes, whereas the *SPP1*$^+$ TAMs and *C1QC*$^+$ TAMs resembled recently described dichotomous functional phenotypes of TAMs in CRC (Zhang et al., 2020). We further confirmed the presence of the *SPP1*$^+$ TAMs in THCA and *FN1*$^+$ TAMs in kidney cancer (KIDNEY) by multi-color IHC (Figure 6D).

The functional phenotypes of macrophages have been reported to exist across an *in vitro* M1/M2 dualistic polarization state (Gordon, 2003). Using signature genes of M1/M2 macrophages, we found the co-expression of both M1 and M2 gene signatures in macrophage subsets from almost all cancer types (Figure S7G; Table S5), consistent with previous studies (Azizi et al., 2018; Zhang et al., 2020). Notably, *SPP1*$^+$ TAMs

**Figure 5. Characterization of cDC2 sub-populations in tumors**

(A) Boxplot showing cell purity for DC subsets. Unpaired two-sided Wilcoxon test.
(B) UMAP plot showing 6 cDC2 subsets.
(C) Violin plots showing gene expression across cDC2 subsets.
(D) Tissue prevalence of each myeloid cluster estimated by *Ro/e* score.
(E) Heatmap showing TF activity for each cDC2 subset.
(F) Pie chart showing the different origins of *LAMP3*$^+$ cDCs predicted by SingleR.
(G) Boxplots comparing the expression of *CXCL9* and *IDO*1 in *CXCL9*$^+$ cDC2 and *CXCL9*$^+$ cDC2-derived *LAMP3*$^+$ cDCs. Unpaired two-sided Wilcoxon test.
(H) Boxplot showing the proportions of C03_cDC2_IL1B (divided by the total cDC2 number) in tumor tissues. Kruskal-Wallis test.
(A)–(G) were plotted by using the integrated data from newly generated datasets in our lab.
See also Figure S6.




*(legend on next page)*

and *C1QC*$^+$ TAMs exhibited higher M2 signatures, while *ISG15*$^+$ TAMs were with higher expression of canonical M1 signatures. Taken together, our findings further demonstrate the limitation of such *in vitro* polarization model and suggest a more complex phenotype of TAMs in the TME.

Previously, we defined dichotomous functional phenotypes of TAMs in CRC (Zhang et al., 2020). Using the angiogenic and phagocytic signatures (Table S5), we assessed the functional phenotypes of each macrophage subset across different cancer types. As expected, the *SPP1*$^+$ TAMs identified in 8 tumor types (BRCA, PAAD, LUNG, CRC, UCEC, NPC, OV-FTC, and THCA) showed preferential expression of genes involved in angiogenesis, whereas *C1QC*$^+$ TAMs exhibited significant higher "phagocytosis scores," which were further validated in LUNG using gene set variation analysis (GSVA) (Figures 6E and S8A). Notably, in tumors with no *SPP1*$^+$ TAMs found, compensatory tumor-enriched macrophages usually existed with a pro-angiogenic signature, including *VCAN*$^+$ TAMs in BRCA and melanoma (MEL), *INHBA*$^+$ TAMs in stomach cancer (STAD) and ESCA, and *FN1*$^+$ TAMs in KIDNEY (Figure 6E). Using clinical data from the TCGA Project, we investigated the relationship of the angiogenesis-associated TAMs with patient prognosis. Overall, the higher expression of *SPP1* was associated with a worse clinical outcome in multiple cancer types (Figure S8B). We also revealed that the higher expression of other marker genes of angiogenesis was linked with a survival disadvantage, such as *FN1* in KIDNEY and *VCAN* in skin cutaneous melanoma (SKCM) (Figure 6F). Thus, angiogenesis-associated macrophages in various tumor types were characterized by diverse markers and were typically associated with poor prognosis.

We next investigated signature genes of those angiogenesis-associated TAMs in corresponding healthy tissues. Here, we used the scRNA-seq data from human cell landscape (HCL) (Han et al., 2020), and applied Scanorama to integrate and normalize the data. Strikingly, we found that these genes always showed much lower expression in the corresponding healthy tissues (Figure S8C), indicating a specific TME-induced expression program in TAMs. Moreover, although a remarkable variation of the proportions of those angiogenesis-associated TAMs in TIMs were observed across various cancer types (Figure S8D), we noticed that these pro-angiogenic macrophage populations tended to be clustered together when quantifying their transcriptomic similarities (Figure 6A), further supporting their shared pro-angiogenic phenotype and functional roles in tumor progression.

We reasoned that their abundance differences across cancer types might influence the diverse responses to immunotherapy. To test this hypothesis, we explored the scRNA-seq data derived from MEL patients treated with immune checkpoint inhibitors (Sade-Feldman et al., 2019). Using the established landscape of TIMs in MEL as a reference, we applied SciBet to predict cell identities for each myeloid cell and identified a consistent pattern of those sub-populations between two datasets (Figure S8E). The angiogenesis-associated *VCAN*$^+$ TAMs were much lower in responders, whereas no significant frequency changes were observed in *C1QC*$^+$ TAMs (Figures 6G and S8F), suggesting the influence of angiogenesis-associated TAMs on diverse responses to immunotherapies.

Using PAGA analysis (Wolf et al., 2019) to decipher the ontogeny origins of different monocyte/macrophage clusters, we compared the developmental potential from CD14$^+$ monocytes to different macrophages clusters. *LYVE1*$^+$ RTMs and *C1QC*$^+$ TAMs showed much weaker connectivity with CD14$^+$ monocytes across multiple tumor types, suggesting their tissue-resident characteristics (Figure S8G). To further explore the diversity of TAMs, we compared our data to another independent HCC scRNA-seq dataset (Sharma et al., 2020). As reported, the tissue-resident *FOLR2*$^+$ TAMs exhibited fetal-liver macrophage (FLM) characteristics and supported an onco-fetal reprogramming of TME in HCC (Sharma et al., 2020). Interestingly, we observed the highest similarities between *LVYE1*$^+$ RTMs and *FOLR2*$^+$ TAMs, as well as FLMs (Figure S8H). Therefore, we hypothesized that the presence of *LVYE1*$^+$ RTMs in adjacent normal tissues may serve as the potential source of the *FOLR2*$^+$ TAMs. To investigate this, we compared the connectivity from *LVYE1*$^+$ RTMs to other macrophages and observed closer connectivity with *C1QC*$^+$ TAMs in our data (Figure 6H). Notably, the *C1QC*$^+$ TAMs also showed a high similarity with *FOLR2*$^+$ TAMs in HCC (Figure S8H), indicating a close relationship between *LVYE1*$^+$ RTMs and *FOLR2*$^+$ TAMs. Taken together, these results indicated the potential of *LVYE1*$^+$ RTMs to develop into onco-fetal *FOLR2*$^+$ TAMs or *C1QC*$^+$ TAMs and their unexpected functions in reprogramming the TME.

## Factors potentially shaping the composition of tumor-infiltrating myeloid cells

The large sample size in this study enabled us to examine multiple factors associated with distinct TIM compositions. First, we

**Figure 6. Characteristics of monocyte and macrophage subsets in various cancer types**

(A) Hierarchical clustering of monocyte and macrophage subsets.

(B) Tissue prevalence of each monocyte and macrophage cluster estimated by *Ro/e* score.

(C) Boxplots comparing the proportion of *SPP1*$^+$ macrophages (divided by the total myeloid cell number) between non-cancer and tumor tissues. Paired two-sided Wilcoxon test.

(D) Representative example of a THCA tumor (top) and a KIDNEY tumor (bottom) stained by IHC. Scale bar, 20 μm. The *FN1*$^+$ TAMs in KIDNEY were characterized with high expression of *CD52*.

(E) Heatmaps showing different expression patterns of angiogenesis- and phagocytosis-associated signature genes among macrophage subsets. *, angiogenesis-associated macrophages.

(F) Kaplan-Meier plots showing worse clinical outcome in KIDNEY patients with the higher expression of *FN1* and in SKCM patients with the higher expression of *VCAN*. +, censored observations; HR, hazard ratio. Cox regression.

(G) Decrease in the frequency of *VCAN*$^+$ macrophages as a proportion of all myeloid cells in the responders of checkpoint inhibitors immunotherapy. Unpaired one-sided Wilcoxon test.

(H) Boxplot showing the connectivity between different macrophage clusters and *LYVE1*$^+$ RTMs. Unpaired two-sided t test.

See also Figures S7 and S8 and Table S5.

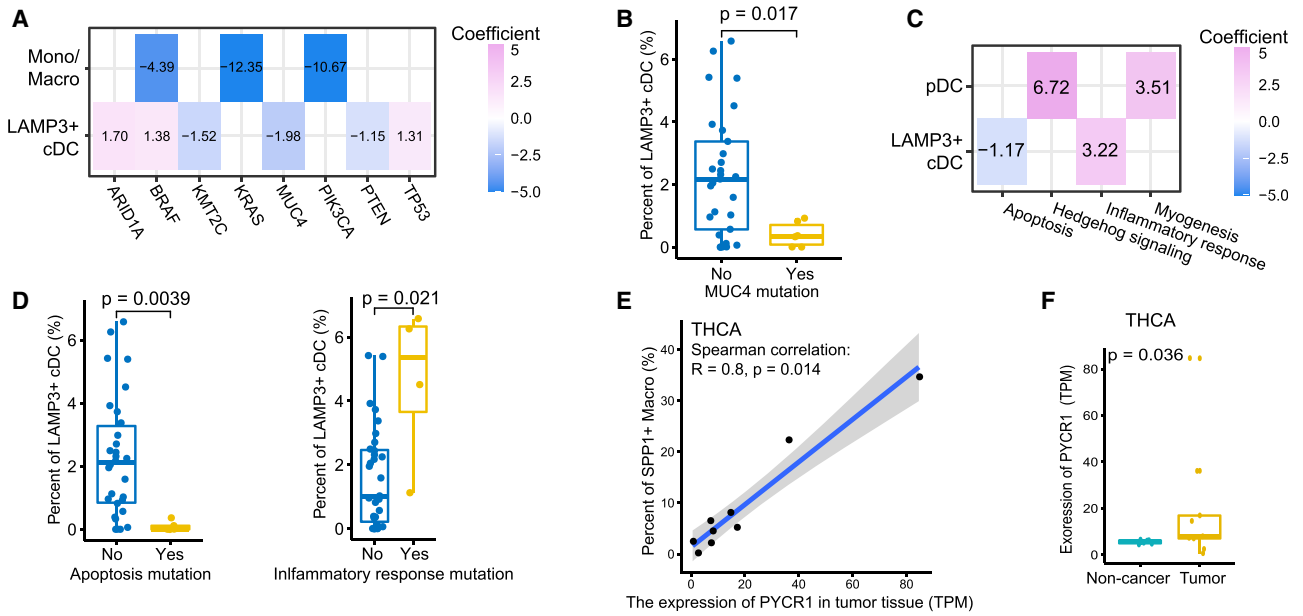



**Figure 7. Factors correlated with the TIM composition**

(A) Heatmap showing the coefficients of genes with somatic mutations significantly associated with the fractions of myeloid subset. Lasso regression.

(B) Boxplot showing the comparison of *LAMP3*$^+$ cDCs fraction between cancer patients with or without mutations in *MUC4*. Unpaired two-sided Wilcoxon test.

(C) Heatmap showing the coefficients of pathways with somatic mutations significantly associated with the fractions of myeloid subset. Lasso regression.

(D) Boxplot showing the comparison of *LAMP3*$^+$ cDCs fraction between cancer patients with or without mutations in apoptosis and inflammatory response pathways. Unpaired two-sided Wilcoxon test.

(E) Scatterplot showing the correlation between *PYCR1* expression and *SPP1*$^+$ TAMs proportion in THCA. Spearman correlation.

(F) Boxplot showing the expression of *PYCR1* in non-cancer and tumor tissues in THCA. Unpaired two-sided Wilcoxon test.

The fraction was calculated by dividing the total myeloid cell number.

See also Table S6.

explored the relationship between TIM compositions and somatic mutation profiles in our pan-cancer collection (STAR methods) and identified several genes with somatic mutations correlated with myeloid cell subset proportions (Figure 7A). Of note, we identified a negative association between the proportion of *LAMP3*$^+$ cDCs and mutations in *MUC4* (Figure 7B), a gene known to be associated with tumor progression via modulating tumor cell apoptosis (Carraway et al., 2009). Furthermore, we detected that mutations in apoptosis and inflammatory response pathways were correlated with proportions of *LAMP3*$^+$ cDCs in tumors (Figures 7C and 7D). Notably, *AXL*, which has been implicated in diverse cancer progression (Rankin and Giaccia, 2016), was among mutated genes in the inflammatory response pathway and was preferentially expressed by cDC2 and *C1QC*$^+$ TAMs.

We also examined the correlation between proportions of TIM subsets and genes upregulated in tumor tissues to reveal possible regulators for TIMs in 5 cancer types with sufficient samples (Table S6; STAR methods). Notably, the *PYCR1* gene expression in THCA was positively correlated with the proportion of *SPP1*$^+$ macrophages (Figure 7E). *PYCR1*, encoding an enzyme involved in the biosynthesis of proline, was upregulated in tumor tissues (Figure 7F) but was not detected in any CD45$^+$ immune cells from THCA (data not shown). Due to the role of proline in maintaining the redox balance of cells and preventing apoptosis, *PYCR1* is emerging as an attractive oncology target and plays important roles in tumor growth (D'Aniello et al., 2020; Loayza-Puch et al., 2016). The relationship between *PYCR1* expression and *SPP1*$^+$ angiogenesis-associated macrophages in THCA may underscore the potential pro-tumorigenic role of *SPP1*$^+$ angiogenesis-associated macrophages.

## DISCUSSION

We collected a large number of myeloid cells from patients diagnosed with one of 15 cancer types to systematically investigate the complexity of TIMs. Surprisingly, distinguished from other cancer types, mast cells in NPC were characterized with high anti-tumor preference, which was supported by their association with better prognosis. Our findings identified specific cancer type with the potential to response to mast cell-targeted immunotherapy, whose possibility and mechanisms should be further explored.

Our pan-cancer scRNA-seq analysis proved that the *LAMP3*$^+$ cDCs were broadly present and extended the conclusion of their diverse origins to all cancer types. Based on the extensive transcriptional analysis, we reason that cDC1-derived and cDC2-derived *LAMP3*$^+$ cDCs are regulated by different ligand-receptor pairs and might have diverse functions. Particularly, in agreement with a recent study that elucidated the dual functions of cDC1-derived *LAMP3*$^+$ cDCs on Tregs and CD8$^+$ T cells (Maier et al., 2020), we also confirmed the complex co-expression

pattern of activation and inhibitory molecules in cDC1-derived *LAMP3*$^+$ cDCs in our pan-cancer analysis. Additionally, during the transition from *CXCL9*$^+$ cDC2s to *LAMP3*$^+$ cDCs, we identified opposite trends of expression for *IDO1* and *CXCL9*, implying the enhanced immunosuppressive ability of cDC2-derived *LAMP3*$^+$ cDCs.

Aside from the reported CD163$^+$CD14$^+$ "pro-inflammatory" cDC2 subset enriched in blood (Dutertre et al., 2019), we identified a tumor-enriched cDC2 subset C03_cDC2_IL1B with high expression of pro-inflammatory mediators, such as *IL1B* and *TNF*. To a certain degree, the *IL1B*$^+$ cDC2s resemble the "pro-inflammatory" cDC2B identified in human spleen, which is also characterized with a high expression of *IL1B* (Brown et al., 2019). Further studies are needed to reveal their functions and differential lineage and migratory relationship with the pro-inflammatory cDC2s in blood and spleen.

Among all tumor-enriched macrophages, *ISG15*$^+$ TAMs upregulated a similar program as C05_cDC2_ISG15, including interferon-induced protein (*ISG15*, *IFIT1*, and *IFITM1*), apoptosis or pyroptosis-mediated death regulators (*TNFSF10*, *CASP1*, and *CASP4*), and chemokine gene *CXCL10*, exhibiting an M1 phenotype bias synergistically induced by type I and type II IFNs. Of note, we highlighted a macrophage sub-population associated with tumor angiogenesis, which was linked with poor prognosis. *SPP1* could be used as the marker gene for the angiogenesis-associated macrophages in 8 tumor types, whereas *SPP1*$^-$ compensatory pro-angiogenic macrophage populations were found in other cancer types. Therefore, immunotherapy targeting those angiogenesis-associated macrophages should take their diverse characteristics in various cancer types into consideration.

In summary, our comprehensive characterization of TIMs revealed their common and specific characteristics among 15 cancer types. To facilitate usage of our data for the wide research community, we developed an interactive web-based tool (http://panmyeloid.cancer-pku.cn) for analyzing and visualizing our single-cell data. Our data can serve as a rich resource to gain deeper understanding of TIMs in various tumors and could provide valuable insights for new therapeutic targeting of myeloid cells.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human specimens
- METHOD DETAILS
  - Single cell collection, sorting, library preparation and sequencing
  - Single-cell RNA-Seq datasets collected in this study
  - Multi-color immunohistochemistry
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Single-cell RNA-seq data processing
  - Dimension reduction and unsupervised clustering
  - Integration of multiple scRNA-seq datasets by Scanorama
  - Evaluating cluster consistency in cancer types with two independent datasets
  - Tissue distribution of clusters
  - Differential expression analysis
  - Similarity analysis of clusters
  - Comparison dendrograms
  - Assessing the heterogeneity of single cell populations
  - Scoring cDCs by using signature genes for each cDC subset
  - Developmental trajectory inference
  - Quantifying the fraction of LAMP3$^+$ cDCs from different origins
  - NicheNet analysis
  - Cell-cell interaction analysis
  - SCENIC analysis
  - Cell identity predicted by SciBet
  - Evaluation of mixability with entropy metric
  - Gene set variation analysis
  - Bulk sequencing and data preprocessing
  - Bulk exome sequencing data analysis
  - Bulk RNA-seq analysis
  - Correlation between myeloid subset composition and other factors
  - TCGA data analysis

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2021.01.010.

## ACKNOWLEDGMENTS

We thank Drs. Nan Wu, Chao Lv, Jinfeng Chen, Zhen Liang (Department of Thoracic Surgery II), Ning Zhang (Department of Urology), Jinfeng Li (Breast Cancer Center), Hong Zheng (Department of gynecologic oncology), and Yang Li (Department of Gastrointestinal Surgery) from Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute and Hanqiang Ouyang and Liang Jiang from Peking University Third Hospital for clinical sample preparation. Part of the experiments were performed on the National Center for Protein Sciences at Peking University. Part of the analysis was performed on the High Performance Computing Platform of the Center for Life Science (Peking University). This project was supported by National Natural Science Foundation of China (81988101, 91742203, 91942307, and 31991171), Beijing Municipal Science and Technology Commission (Z201100005320014), and Beijing Advanced Innovation Center for Genomics. H.O. and L.J. were supported by AOSpine 2019 Discovery and Innovation Award (AOSDIA2019-026) and Beijing Natural Science Foundation (7204327).

## AUTHOR CONTRIBUTIONS

Z.Z., J.J., and Z.B. designed experiments. S.C., Z.L., Y.Y., L.Z., X.R., and X.W. performed scRNA-seq data analysis. Y.G., H.O., P.D., L.J., B.Z., Y.Y., and B.X. collected clinical samples. S.Q. performed bulk exome sequencing and RNA-seq data analysis. X.H. coordinated patient sample collections. R.G. performed the experiments. Z.L. and S.C. constructed the website. S.C., Z.L.,

and Z.Z. wrote the manuscript, with all authors contributing to writing and providing feedback.

## DECLARATION OF INTERESTS

Z.Z. is a founder of Analytical Biosciences and also a scientific advisor for InnoCare Pharma and ArsenalBio. Other authors declare no conflicts of interest.

Received: July 9, 2020
Revised: November 16, 2020
Accepted: January 7, 2021
Published: February 4, 2021

## REFERENCES

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172.

Azizi, E., Carr, A.J., Plitas, G., Cornish, A.E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., et al. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell *174*, 1293–1308.e36.

Barnes, P.J. (1999). Effect of beta-agonists on inflammatory cells. J. Allergy Clin. Immunol. *104*, S10–S17.

Binnewies, M., Mujal, A.M., Pollack, J.L., Combes, A.J., Hardison, E.A., Barry, K.C., Tsui, J., Ruhland, M.K., Kersten, K., Abushawish, M.A., et al. (2019). Unleashing Type-2 Dendritic Cells to Drive Protective Antitumor $CD4^+$ T Cell Immunity. Cell *177*, 556–571.e16.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nat. Methods *17*, 159–162.

Brown, C.C., Gudjonson, H., Pritykin, Y., Deep, D., Lavallée, V.P., Mendoza, A., Fromme, R., Mazutis, L., Ariyan, C., Leslie, C., et al. (2019). Transcriptional Basis of Mouse and Human Dendritic Cell Heterogeneity. Cell *179*, 846–863.e24.

Burr, M.L., Sparbier, C.E., Chan, Y.C., Williamson, J.C., Woods, K., Beavis, P.A., Lam, E.Y.N., Henderson, M.A., Bell, C.C., Stolzenburg, S., et al. (2017). CMTM6 maintains the expression of PD-L1 and regulates anti-tumour immunity. Nature *549*, 101–105.

Carraway, K.L., Theodoropoulos, G., Kozloski, G.A., and Carothers Carraway, C.A. (2009). Muc4/MUC4 functions and regulation in cancer. Future Oncol. *5*, 1631–1640.

Castellino, F., Huang, A.Y., Altan-Bonnet, G., Stoll, S., Scheinecker, C., and Germain, R.N. (2006). Chemokines enhance immunity by guiding naive CD8+ T cells to sites of CD4+ T cell-dendritic cell interaction. Nature *440*, 890–895.

Chakarov, S., Lim, H.Y., Tan, L., Lim, S.Y., See, P., Lum, J., Zhang, X.M., Foo, S., Nakamizo, S., Duan, K., et al. (2019). Two distinct interstitial macrophage populations coexist across tissues in specific subtissular niches. Science *363*, eaau0964.

Crotty, S., Johnston, R.J., and Schoenberger, S.P. (2010). Effectors and memories: Bcl-6 and Blimp-1 in T and B lymphocyte differentiation. Nat. Immunol. *11*, 114–120.

D'Aniello, C., Patriarca, E.J., Phang, J.M., and Minchiotti, G. (2020). Proline Metabolism in Tumor Growth and Metastatic Progression. Front. Oncol. *10*, 776.

DeNardo, D.G., and Ruffell, B. (2019). Macrophages as regulators of tumour immunity and immunotherapy. Nat. Rev. Immunol. *19*, 369–382.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dudeck, J., Ghouse, S.M., Lehmann, C.H., Hoppe, A., Schubert, N., Nedospasov, S.A., Dudziak, D., and Dudeck, A. (2015). Mast-Cell-Derived TNF Amplifies CD8(+) Dendritic Cell Functionality and CD8(+) T Cell Priming. Cell Rep. *13*, 399–411.

Dutertre, C.A., Becht, E., Irac, S.E., Khalilnezhad, A., Narang, V., Khalilnezhad, S., Ng, P.Y., van den Hoogen, L.L., Leong, J.Y., Lee, B., et al. (2019). Single-Cell Analysis of Human Mononuclear Phagocytes Reveals Subset-Defining Markers and Identifies Circulating Inflammatory Dendritic Cells. Immunity *51*, 573–589.e8.

Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat. Protoc. *15*, 1484–1506.

Engblom, C., Pfirschke, C., and Pittet, M.J. (2016). The role of myeloid cells in cancer therapies. Nat. Rev. Cancer *16*, 447–462.

Ginhoux, F., Schultze, J.L., Murray, P.J., Ochando, J., and Biswas, S.K. (2016). New insights into the multidimensional concept of macrophage ontogeny, activation and function. Nat. Immunol. *17*, 34–40.

Gordon, S. (2003). Alternative activation of macrophages. Nat. Rev. Immunol. *3*, 23–35.

Gordon, S., Plüddemann, A., and Martinez Estrada, F. (2014). Macrophage heterogeneity in tissues: phenotypic diversity and functions. Immunol. Rev. *262*, 36–55.

Gosselin, D., Link, V.M., Romanoski, C.E., Fonseca, G.J., Eichenfield, D.Z., Spann, N.J., Stender, J.D., Chun, H.B., Garner, H., Geissmann, F., and Glass, C.K. (2014). Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. Cell *159*, 1327–1340.

Gulati, G.S., Sikandar, S.S., Wesche, D.J., Manjunath, A., Bharadwaj, A., Berger, M.J., Ilagan, F., Kuo, A.H., Hsieh, R.W., Cai, S., et al. (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. Science *367*, 405–411.

Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., Kang, B., Liu, Z., Jin, L., Xing, R., et al. (2018). Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat. Med. *24*, 978–985.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hashimoto-Hill, S., Friesen, L., Park, S., Im, S., Kaplan, M.H., and Kim, C.H. (2018). RARα supports the development of Langerhans cells and langerin-expressing conventional dendritic cells. Nat. Commun. *9*, 3896.

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685–691.

Ishida, T., and Ueda, R. (2006). CCR4 as a novel molecular target for immunotherapy of cancer. Cancer Sci. *97*, 1139–1146.

Jang, J.E., Hajdu, C.H., Liot, C., Miller, G., Dustin, M.L., and Bar-Sagi, D. (2017). Crosstalk between Regulatory T Cells and Tumor-Associated Dendritic Cells Negates Anti-tumor Immunity in Pancreatic Cancer. Cell Rep. *20*, 558–571.

Kenkel, J.A., Tseng, W.W., Davidson, M.G., Tolentino, L.L., Choi, O., Bhattacharya, N., Seeley, E.S., Winer, D.A., Reticker-Flynn, N.E., and Engleman, E.G. (2017). An Immunosuppressive Dendritic Cell Subset Accumulates at Secondary Sites and Promotes Metastasis in Pancreatic Cancer. Cancer Res. *77*, 4158–4170.

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S., and Amit, I. (2014). Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell *159*, 1312–1326.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., Ren, X., and Zhang, Z. (2020). SciBet as a portable and fast single cell type identifier. Nat. Commun. *11*, 1818.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al.; Cancer Genome Atlas Research Network (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell *173*, 400–416.e11.

Liu, B., Li, C., Li, Z., Wang, D., Ren, X., and Zhang, Z. (2020). An entropy-based metric for assessing the purity of single cell populations. Nat. Commun. *11*, 3155.

Loayza-Puch, F., Rooijers, K., Buil, L.C., Zijlstra, J., Oude Vrielink, J.F., Lopes, R., Ugalde, A.P., van Breugel, P., Hofland, I., Wesseling, J., et al. (2016). Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. Nature *530*, 490–494.

Maciel, T.T., Moura, I.C., and Hermine, O. (2015). The role of mast cells in cancers. F1000Prime Rep. *7*, 09.

Maeda, M., Yanagawa, Y., Iwabuchi, K., Minami, K., Nakamaru, Y., Takagi, D., Fukuda, S., and Onoé, K. (2007). IL-21 enhances dendritic cell ability to indu,ce interferon-gamma production by natural killer T cells. Immunobiology *212*, 537–547.

Maier, B., Leader, A.M., Chen, S.T., Tung, N., Chang, C., LeBerichel, J., Chudnovskiy, A., Maskey, S., Walker, L., Finnigan, J.P., et al. (2020). A conserved dendritic-cell regulatory program limits antitumour immunity. Nature *580*, 257–262.

Marone, G., Varricchi, G., Loffredo, S., and Granata, F. (2016). Mast cells and basophils in inflammatory and tumor angiogenesis and lymphangiogenesis. Eur. J. Pharmacol. *778*, 146–151.

Maynard, A., McCoach, C.E., Rotow, J.K., Harris, L., Haderk, F., Kerr, D.L., Yu, E.A., Schenk, E.L., Tan, W., Zee, A., et al. (2020). Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. Cell *182*, 1232–1251.e22.

Mellman, I. (2013). Dendritic cells: master regulators of the immune response. Cancer Immunol. Res. *1*, 145–149.

Merad, M., Ginhoux, F., and Collin, M. (2008). Origin, homeostasis and function of Langerhans cells and other langerin-expressing dendritic cells. Nat. Rev. Immunol. *8*, 935–947.

Mezzadra, R., Sun, C., Jae, L.T., Gomez-Eerland, R., de Vries, E., Wu, W., Logtenberg, M.E.W., Slagter, M., Rozeman, E.A., Hofland, I., et al. (2017). Identification of CMTM6 and CMTM4 as PD-L1 protein regulators. Nature *549*, 106–110.

Nakamura, K., and Smyth, M.J. (2020). Myeloid immunosuppression and immune checkpoints in the tumor microenvironment. Cell. Mol. Immunol. *17*, 1–12.

Peng, J., Sun, B.F., Chen, C.Y., Zhou, J.Y., Chen, Y.S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G.S., et al. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. *29*, 725–738.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. Nat. Methods *14*, 309–315.

Rankin, E.B., and Giaccia, A.J. (2016). The Receptor Tyrosine Kinase AXL in Cancer Progression. Cancers (Basel) *8*, 103.

Ren, X., Kang, B., and Zhang, Z. (2018). Understanding tumor ecosystems by single-cell sequencing: promises and limitations. Genome Biol. *19*, 211.

Ribatti, D. (2016). Mast cells as therapeutic target in cancer. Eur. J. Pharmacol. *778*, 152–157.

Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., Ray, J.P., de Boer, C.G., Jenkins, R.W., Lieb, D.J., Chen, J.H., Frederick, D.T., Barzily-Rokni, M., et al. (2019). Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. Cell *176*, 404.

Saikh, K.U., Khan, A.S., Kissner, T., and Ulrich, R.G. (2001). IL-15-induced conversion of monocytes to mature dendritic cells. Clin. Exp. Immunol. *126*, 447–455.

Salmon, H., Idoyaga, J., Rahman, A., Leboeuf, M., Remark, R., Jordan, S., Casanova-Acebes, M., Khudoynazarova, M., Agudo, J., Tung, N., et al. (2016). Expansion and Activation of CD103(+) Dendritic Cell Progenitors at the Tumor Site Enhances Tumor Responses to Therapeutic PD-L1 and BRAF Inhibition. Immunity *44*, 924–938.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811–1817.

Schmid, M.C., and Varner, J.A. (2010). Myeloid cells in the tumor microenvironment: modulation of tumor angiogenesis and tumor inflammation. J. Oncol. *2010*, 201026.

Scott, C.L., and Omilusik, K.D. (2019). ZEBs: Novel Players in Immune Cell Development and Function. Trends Immunol. *40*, 431–446.

Sharma, A., Seow, J.J.W., Dutertre, C.A., Pai, R., Blériot, C., Mishra, A., Wong, R.M.M., Singh, G.S.N., Sudhagar, S., Khalilnezhad, S., et al. (2020). Onco-fetal Reprogramming of Endothelial Cells Drives Immunosuppressive Macrophages in Hepatocellular Carcinoma. Cell *183*, 377–394.e21.

Simon, T., and Bromberg, J.S. (2016). BTLA$^{+}$ Dendritic Cells: The Regulatory T Cell Force Awakens. Immunity *45*, 956–958.

Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705.

Tokunaga, R., Zhang, W., Naseem, M., Puccini, A., Berger, M.D., Soni, S., McSkane, M., Baba, H., and Lenz, H.J. (2018). CXCL9, CXCL10, CXCL11/CXCR3 axis for immune activation - A target for novel cancer therapy. Cancer Treat. Rev. *63*, 40–47.

Trinchieri, G. (2003). Interleukin-12 and the regulation of innate resistance and adaptive immunity. Nat. Rev. Immunol. *3*, 133–146.

Villani, A.C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science *356*, eaah4573.

Vivian, J., Rao, A.A., Nothaft, F.A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A.D., Musselman-Brown, A., et al. (2017). Toil enables reproducible, open source, big biomedical data analyses. Nat. Biotechnol. *35*, 314–316.

Vogel, D.Y., Glim, J.E., Stavenuiter, A.W., Breur, M., Heijnen, P., Amor, S., Dijkstra, C.D., and Beelen, R.H. (2014). Human macrophage polarization in vitro: maturation and activation methods compared. Immunobiology *219*, 695–703.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Woidacki, K., Jensen, F., and Zenclussen, A.C. (2013). Mast cells as novel mediators of reproductive processes. Front. Immunol. *4*, 29.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59.

Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. *8*, 281–291.e9.

Wu, H., Gong, J., and Liu, Y. (2018). Indoleamine 2, 3-dioxygenase regulation of immune response (Review). Mol. Med. Rep. *17*, 4867–4873.

Yang, J., Zhang, L., Yu, C., Yang, X.F., and Wang, H. (2014). Monocyte and macrophage differentiation: circulation inflammatory monocyte as biomarker for inflammatory diseases. Biomark. Res. *2*, 1.

Young, M.D., Mitchell, T.J., Vieira Braga, F.A., Tran, M.G.B., Stewart, B.J., Ferdinand, J.R., Collord, G., Botting, R.A., Popescu, D.M., Loudon, K.W., et al. (2018). Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science *361*, 594–599.

Zhang, X., Gu, J., Yu, F.S., Zhou, L., and Mi, Q.S. (2016). TGF-β1-induced transcription factor networks in Langerhans cell development and maintenance. Allergy *71*, 758–764.

Zhang, L., Yu, X., Zheng, L., Zhang, Y., Li, Y., Fang, Q., Gao, R., Kang, B., Zhang, Q., Huang, J.Y., et al. (2018). Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature *564*, 268–272.

Zhang, Q., He, Y., Luo, N., Patel, S.J., Han, Y., Gao, R., Modak, M., Carotta, S., Haslinger, C., Kind, D., et al. (2019). Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. Cell *179*, 829–845.e20.

Zhang, L., Li, Z., Skrzypczynska, K.M., Fang, Q., Zhang, W., O'Brien, S.A., He, Y., Wang, L., Zhang, Q., Kim, A., et al. (2020). Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell *181*, 442–459.e29.

Zilionis, R., Engblom, C., Pfirschke, C., Savova, V., Zemmour, D., Saatcioglu, H.D., Krishnan, I., Maroni, G., Meyerovitz, C.V., Kerwin, C.M., et al. (2019). Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. Immunity *50*, 1317–1334.e10.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Antibodies** | | |
| 7-AAD Viability Staining Solution (FACS) | eBioscienc | Cat# 00-6993-50 |
| Anti-Human CD45 eFluor 450 (FACS) | eBioscienc | Cat# 48-0459-42 |
| Anti-Human TPSAB1 | Abcam | Cat# 2378 |
| Anti-Human TNF | Abcam | Cat# 220210 |
| Anti-Human VEGFA | Abcam | Cat# 213244 |
| Anti-Human CD68 | Abcam | Cat# 213363 |
| Anti-Human SPP1 | Abcam | Cat# 214050 |
| Anti-Human CD52 | Abcam | Cat# 234412 |
| Anti-Human CD4 | Abcam | Cat# 133616 |
| Anti-Human CD8 | Abcam | Cat# 14147 |
| Anti-Human DC-LAMP | Novus Biologicals | Cat# 1010E1.01 |
| Anti-Human FOXP3 | Abcam | Cat# 22510 |
| **Biological samples** | | |
| Human primary tumor samples | Peking University Cancer Hospital/Peking University Third Hospital | See Table S1 for details |
| Human peripheral blood from tumor patients | Peking University Cancer Hospital/Peking University Third Hospital | See Table S1 for details |
| Human non-cancer tissues from tumor patients | Peking University Cancer Hospital/Peking University Third Hospital | See Table S1 for details |
| **Deposited data** | | |
| Data files for human scRNA-seq | This paper | GEO: GSE154763 |
| Human lung cancer treatment-related scRNA-seq dataset | Maynard et al., 2020 | https://drive.google.com/drive/folders/1nONsp9VuhmPzuDvMet0i8x26eV9r5lkT?usp=sharing |
| Human cell landscape (HCL) dataset | Han et al., 2020 | https://figshare.com/articles/dataset/HCL_DGE_Data/7235471 |
| Human melanoma treatment-related dataset | Sade-Feldman et al., 2019 | GEO: GSE120575 |
| **Software and algorithms** | | |
| Cell Ranger 3.0.0 | 10x Genomics | https://10xgenomics.com/ |
| Scanpy 1.4.3 | Wolf et al., 2018 | https://github.com/theislab/scanpy |
| Scanorama 1.4 | Hie et al., 2019 | https://github.com/brianhie/scanorama |
| ROUGE 1.0 | Liu et al., 2020 | https://github.com/PaulingLiu/ROGUE |
| SingleR 1.0.5 | Aran et al., 2019 | https://github.com/dviraran/SingleR |
| SCENIC 1.1.2-2 | Aibar et al., 2017 | https://github.com/aertslab/SCENIC |
| Monocle 2.14.0 | Qiu et al., 2017 | https://git.bioconductor.org/packages/monocle |
| CytoTRACE 0.3.2 | Gulati et al., 2020 | https://cytotrace.stanford.edu/ |
| CellPhoneDB 2.1.4 | Efremova et al., 2020 | https://www.cellphonedb.org/ |
| NicheNet 0.1.0 | Browaeys et al., 2020 | https://github.com/saeyslab/nichenetr |
| SciBet 1.0 | Li et al., 2020 | https://github.com/PaulingLiu/scibet |
| BWA 0.7.17 | Li and Durbin, 2010 | https://github.com/lh3/bwa |
| Picard 2.18.9 | N/A | http://broadinstitute.github.io/picard/ |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| GATK 3.8 | DePristo et al., 2011 | https://github.com/broadinstitute/gatk |
| Strelka 1.0.14 | Saunders et al., 2012 | ftp://strelka@ftp.illumina.com/v1-branch/v1.0.14/strelka_workflow-1.0.14.tar.gz |
| ANNOVAR 20180416 | Wang et al., 2010 | https://doc-openbio.readthedocs.io/projects/annovar/en/latest/ |
| RSEM 1.3.1 | Li and Dewey, 2011 | https://github.com/deweylab/RSEM |
| Other | | |
| Interactive explorer of human myeloid cells | This paper | http://panmyeloid.cancer-pku.cn |

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Zemin Zhang (zemin@pku.edu.cn).

### Materials availability

This study did not generate new unique reagents.

### Data and code availability

The expression data can be obtained from Gene Expression Omnibus (GEO: GSE154763). The human cell landscape (HCL) dataset (Han et al., 2020) and melanoma treatment-related dataset (Sade-Feldman et al., 2019) were available from https://figshare.com/articles/dataset/HCL_DGE_Data/7235471 and GEO: GSE120575, respectively. Analysis and visualization of the scRNA-seq datasets in this study can also be performed at http://panmyeloid.cancer-pku.cn. Our analysis code has been uploaded into the github (https://github.com/Sijin-ZhangLab/PanMyeloid). The data that support the findings of this study are also available from the corresponding author upon request.

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Human specimens

Forty-eight cancer patients, including thirty-five females and thirteen males, were enrolled and pathologically diagnosed with different cancer types. Their ages ranged from 20 to 80. The stages of these patients were classified according to the guidance of AJCC version 8. Written informed consent was provided by all patients. This study was approved by the Research and Ethical Committee of Peking University Cancer Hospital and complied with all relevant ethical regulations. Fresh tumor samples were surgically resected from each patient, with adjacent normal tissues or peripheral blood as controls. None of patients were treated with chemotherapy or radiation prior to tumor resection. The detailed information is summarized in Table S1.

## METHOD DETAILS

### Single cell collection, sorting, library preparation and sequencing

Single cells were collected from tumor and adjacent non-cancer tissues as described previously (Zhang et al., 2020). Briefly, tumors and adjacent non-cancer tissues were cut into approximately 1-2 $mm^3$ pieces in the RPMI-1640 medium (GIBCO) with 10% fetal bovine serum (FBS, GIBCO), and enzymatically digested with gentleMACS (Miltenyi) for 60 min on a rotor at 37°C, according to manufacturer's instruction. The dissociated cells were subsequently passed through a 100 μm SmartStrainer and centrifuged at 400 g for 5 min. After the supernatant was removed, the pelleted cells were suspended in red blood cell lysis buffer (TIANDZ) and incubated on ice for 1-2 min to lyse red blood cells. After washing twice with 1 × PBS (GIBCO), the cell pellets were re-suspended in sorting buffer (PBS supplemented with 1% FBS).

Single cell suspensions were stained with antibodies against CD45 and 7AAD for FACS sorting, performed on a BD Aria SORP instrument. Based on FACS analysis, single cells were sorted into 1.5 mL tubes (Eppendorf) and counted manually under the microscope. The concentration of single cell suspensions was adjusted to 500-1200 cells/ul. Cells were loaded between 7,000 and 15,000 cells/chip position using the 10x Chromium Single cell 5′ Library, Gel Bead & Multiplex Kit and Chip Kit (10x Genomics, V3 barcoding chemistry) according to the manufacturer's instructions. All the subsequent steps were performed following the standard manufacturer's protocols. Purified libraries were analyzed by an Illumina Hiseq X Ten sequencer with 150-bp paired-end reads.

### Single-cell RNA-Seq datasets collected in this study

We obtained scRNA-seq data on myeloid cells in 380 samples from 210 patients diagnosed with one of the 15 common cancer types (Figure 1A; Table S1). Data from 128 patients (227 samples) were obtained from published studies, including those previously generated by us. To supplement the publicly available data and cover the cancer types that have not been adequately studied, we collected an additional cohort of 82 treatment-naive patients of 10 cancer types (including two unpublished datasets generated previously) (Figure 1A; Table S1), and obtained the scRNA-seq data for 153 samples derived from these patients using the 10X Genomics platform.

### Multi-color immunohistochemistry

Human tissue specimens were provided by Peking University Cancer Hospital & Institute. The specimens were collected within 30 min after the tumor resection and fixed in formalin for 48 hr. Dehydration and embedding in paraffin was performed following routine methods. Paraffin blocks were cut into 5 μm slides and adhered on the slides glass. Then the paraffin sections were placed in the 70°C paraffin oven for 1 hr before deparaffinized in xylene and then rehydrated in 100%, 90%, 70% alcohol successively. Antigen was retrieved by critic acid buffer (pH 6.0) in the 95°C water bath for 20 min. Endogenous peroxidase was inactivated by incubation in 3% $H_2O_2$ for 15 min. Following a preincubation with 10% normal goat serum to block nonspecific sites for 30 min, the sections were incubated with primary antibodies in a humidified chamber at 4°C overnight. The primary antibodies and IHC metrics used in the validation of two mast cell subtypes were: mouse anti-TPSAB1 (Abcam, Cat# 2378), rabbit anti-TNF (Abcam, Cat# 220210) and rabbit anti-VEGFA (Abcam, Cat# 213244). The primary antibodies used in the validation of *LAMP3*$^+$ cDCs were: rabbit anti-human CD4 (Abcam, Cat# 133616), mouse anti-human CD8 (Abcam, Cat# 14147), mouse anti-human FOXP3 (Abcam, Cat# 22510), rat anti-human DC_LAMP (Novus Biologicals, Cat# 1010E1.01). The primary antibodies and IHC metrics used in the validation of the *SPP1*$^+$ TAMs and *FN1*$^+$ TAMs were: rabbit anti-CD68 (Abcam, Cat# ab213363), rabbit anti-SPP1 (Abcam, Cat# 214050), and rabbit anti-CD52 (Abcam, Cat# 234412). After the sections were washed with PBS twice for 5 min, the antigenic binding sites were visualized using the GTVisionTMIIDetection System/Mo&Rb according to the manufacturer's protocol.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Single-cell RNA-seq data processing

The newly generated scRNA-seq data from 10X Genomics were aligned and quantified using the Cell Ranger Single-Cell Software Suite against the GRCh38 human reference genome. The preliminary filtered data generated from Cell Ranger were used for downstream analysis. Further quality control was applied to cells based on three metrics step by step, including the total UMI count, number of detected genes and proportion of mitochondrial gene count per cell. Specifically, cells with less than 2000 UMI count and 500 detected genes were filtered, as well as cells with more than 10% mitochondrial gene count. To remove potential doublets, cells with UMI count above 40,000 and detected genes above 5,000 are also filtered out. Notably, we removed the *CD3E*$^+$ myeloid cluster with a large fraction of potential doublets predicted by Scrublet (Wolock et al., 2019), which expressed both myeloid gene signature (*CST3*) and T cell signature (*CD3E*).

For other collected scRNA-seq datasets, we applied the same filtering steps to 10X Genomics datasets (LUNG, KIDNEY, STAD, CRC, HCC, NPC and PAAD). For inDrop datasets (BRCA), we removed cells with less than 300 total count and cells with more than 20% mitochondrial gene count. For cells generated by MARS-Seq (MEL), we removed cells with less than 300 total count and cells with more than 20% ERCC gene count.

After quality control, we applied the library-size correction method to normalize the raw count by using *normalize_total* function in Scanpy (Wolf et al., 2018). Then the logarithmized normalized count matrix was used for the downstream analysis.

### Dimension reduction and unsupervised clustering

Single cell data were processed for dimension reduction and unsupervised clustering by following the workflow in Scanpy (Wolf et al., 2018). In brief, 2,000 highly-variable genes were selected for downstream analysis by using *scanpy.pp.highly_variable_genes* function with parameter "*n_top_genes=2000*." Then, effects of the total count per cell, the percentage of mitochondrial gene count and the percentage of count for heat shock protein associated genes (HSP) were regressed out by using *scanpy.pp.regress_out* function. A principal component analysis (PCA) matrix with 100 components were calculated to reveal the main axes of variation and denoise the data by using *scanpy.tl.pca* function with parameter "*svd_solver='arpack', n_comps=100*." To remove the batch effects from different donors, we applied *bbknn* algorithm with parameter "*batch_key='patient', n_pcs=100*" to obtain a batch-corrected space. For visualization, the dimensionality of each dataset was further reduced using Uniform Manifold Approximation and Projection (UMAP) implemented in *scanpy.tl.umap* function with default parameters. To cluster single cells by their expression profiles, we used an unsupervised graph-based clustering algorithm called *Leiden* (we used different resolutions range from 0.6 to 1 for different datasets). The cluster-specific marker genes were identified by using the *scanpy.tl.rank_genes_groups* function with default parameters. For each dataset, we first performed unsupervised clustering and characterized myeloid cell clusters with high expression of *PTPRC* and *CST3*, and a second-round of dimension reduction and unsupervised clustering were performed to characterize the subsets of myeloid cells following the above-described strategy.

### Integration of multiple scRNA-seq datasets by Scanorama

We run two rounds of Scanorama (Hie et al., 2019), an algorithm that could identify and merge shared cell types among multiple datasets, to remove the batch effects within scRNA-seq datasets of 15 cancer types. First, we applied Scanorama to datasets generated from 3′ library and 5′ library from 10x Genomics to remove the batch effects attribute to these two protocols. Then, a second-round of Scanorama was applied to remove the batch effects resulting from the diverse platforms, including 10x Genomics, MARS-Seq and inDrop. The results from Scanorama integration and batch correction were then used as input data for scRNA-seq analysis, including highly-variable gene identification and dimension reduction.

### Evaluating cluster consistency in cancer types with two independent datasets

For the datasets of four cancer types (LUNG, KIDNEY, PAAD and STAD), each comprising two independent studies, we first performed graph-based clustering on each study separately (Figures S1C and S2B), and then applied Scanorama (Hie et al., 2019) to integrate each of the four cancer types. Cells annotated as the same subsets were mixed together (Figures 1G and S2C). Moreover, we trained a logistical regression model (Young et al., 2018; Zhang et al., 2020) to quantify the similarities between clusters from different studies (Logistic regression model, below). As expected, the clusters showed remarkable concordance across different datasets (Figure S2D). Thus, the consistent cluster composition identified from different datasets further verified the rationality of the cluster identification.

### Tissue distribution of clusters

We calculated the ratio of observed to expected cell numbers (*Ro/e*) for each cluster in different tissues to quantify the tissue preference of each cluster (Guo et al., 2018; Zhang et al., 2018). The expected cell numbers for each combination of cell clusters and tissues were obtained from the chi-square test. One cluster was identified as being enriched in a specific tissue if $Ro/e > 1$.

### Differential expression analysis

To identify differentially expressed genes between two groups of clusters, we used *wilcox.test* in R to evaluate the significance of each gene, with multiple hypothesis correction using the Benjamini-Hochberg procedure. Genes with adjusted *P*-value less than 0.05 were considered as differentially expressed genes. In addition, the log2 fold change (log2FC) for each gene was calculated by subtracting the log2 transformed mean count in each group.

### Similarity analysis of clusters

#### *Logistic regression model*

Here we used two different methods to evaluate the cluster similarities. To measure the similarities of clusters identified from different datasets (Figures S2D and S4C), we trained a logistic regression model using elastic net regularization (Young et al., 2018; Zhang et al., 2020). Briefly, the R function *cv.glmnet* from the glmnet package was used to fit a series of N binomial logistic regression models with parameters alpha = 0.99. The N represented the number of clusters in the training data. To eliminate possible bias resulting from different cluster sizes in the training dataset, we downsampled each cluster to the minimum size of all clusters. The sample size was set as 50 when the minimum size was less than 50. The offset for each model was calculated with $\log f/(1-f)$), where f is the fraction of cells in the cluster being trained.

In each case, a 10-fold cross validation was performed. We used these models to calculate a predicted logit of each cell in the test data for each cluster in the training data with an offset of 0. Predicted logits were then averaged within each cluster and converted to probabilities for visualization, indicating the similarity of clusters from the test data to those from the training data.

### Comparison dendrograms

For an unsupervised comparison of the sub-populations identified from multiple cancer types, the following steps were performed. Here, we used the batch-corrected expression value from Scanorama.

(1) Identify a set of highly variable genes across different sub-populations.
(2) Calculate the mean expression of genes in each cluster.
(3) For hierarchical clustering, the distance defined as (1-Pearson correlation coefficient)/2 was used.

For major myeloid lineages comparison across multiple cancer types, we used the top 2,000 highly variable genes. For other sub-populations comparison, top 1,000 highly variable genes were used.

### Assessing the heterogeneity of single cell populations

We used ROGUE (Liu et al., 2020), an entropy-based universal metric for assessing the purity of single cell population, to compare the heterogeneity of major myeloid lineages and DC sub-populations in our study. The ROGUE index has been scaled to between zero and one. One represents a completely pure subtype with no significant genes and zero represents the most heterogeneous state of a population.

### Scoring cDCs by using signature genes for each cDC subset

To trace the origins of *LAMP3*$^+$ cDCs, we used the signature genes generated from each cDC subset to score all cDCs and the stratification of cDCs would give a reliable prediction for their origins (Maier et al., 2020). Briefly, signature genes for cDC1 and cDC2 were defined as genes with an absolute log2FC more than 0.3 and adjusted *P*-value less than 0.05 when comparing the transcriptome of cDC1s to cDC2s. Signature genes for *LAMP3*$^+$ cDCs were defined as genes with a log2FC more than 0.5 and adjusted *P*-value less than 0.05 when comparing the transcriptome of *LAMP3*$^+$ cDCs to both cDC1s and cDC2s. Next, we calculated three scores for all cDCs to measure their characteristics of cDC1, cDC2 and *LAMP3*$^+$ cDCs. Each score was calculated as the fraction of RNA in a cell belonging to signature genes. To compare cDC1 and cDC2 scores for each cell, the signature scores in each cell were defined as:

$$\sum_{\text{signature genes}} E_{i,j} \Big/ \left( \sum_{\text{all genes}} E_{i,j} \times N \right)$$

where $E_{i,j}$ referred to the expression level of gene j in cell i and N was the number of signature genes. Finally, all scores plotted in Figure S5B were log10 transformed.

### Developmental trajectory inference

#### *Monocle*

To characterize the developmental origins of *LAMP3*$^+$ cDCs, we applied the Monocle (version 2.14.0) algorithm (Qiu et al., 2017) with the top 400 signature genes calculated by *differentialGeneTest* function. The cDC cell differentiation trajectory was inferred with the default parameters of Monocle after dimension reduction and cell ordering.

#### *CytoTRACE*

We also performed CytoTRACE analysis with default parameter (Gulati et al., 2020), an algorithm that predict differentiation states from scRNA-seq data based on the simple yet robust observation that transcriptional diversity decreases during differentiation, to complement the trajectory analysis from Monocle.

#### *PAGA*

To assess the most likely trajectories of cell progression among myeloid cells in each cancer type, monocyte and macrophage subpopulations were subjected to the partition-based graph abstraction method PAGA (Wolf et al., 2018), a part of the single-cell analysis package Scanpy (Wolf et al., 2018). The computations were carried out using default parameters. The edge connectivity between each subpopulation node for all edges are further compared by using unpaired two-sided Student's t test.

### Quantifying the fraction of LAMP3$^+$ cDCs from different origins

To quantify the fraction of *LAMP3*$^+$ cDCs from different origins, we used SingleR (Aran et al., 2019), a predictor mainly based on Spearman coefficients of single-cell gene expression with each of the samples in the reference dataset. Here, we used the transcriptomes of cDC1s and cDC2s as a reference to predict the origin of *LAMP3*$^+$ cDCs in each dataset. Besides, we also used cDC2 subsets as a reference to study the relationship between cDC2 subsets and *LAMP3*$^+$ cDCs.

### NicheNet analysis

NicheNet (Browaeys et al., 2020) is a powerful tool that predicts ligands driving the transcriptomic changes of target cells. To identify potential ligands that drive the unique phenotype of *TNF*$^+$ mast cells, we first calculated the differently expressed genes (DEGs) between *TNF*$^+$ and *TNF*$^-$ mast cells. Then, genes with log2FC > 0.2 and adjusted *P*-value < 0.05 were used as gene sets of interest. All expressed genes in *TNF*$^+$ mast cells were used as background of genes. Genes were considered as expressed when they have non-zero values in at least 10% of the cells in a cell type. In general, users could use the expression profile of specific sender cells and target cells to construct the expressed ligand-receptor interactions and narrow down the number of ligands to be evaluated. Here, we only used the expressed receptors in *TNF*$^+$ mast cells to construct the expressed ligand-receptor interactions and calculate the ligand activity.

We also applied NicheNet analysis to uncover the diverse external stimulus for *LAMP3*$^+$ cDCs from different origins. We first calculated two groups of differently expressed genes (DEGs): cDC1-derived *LAMP3*$^+$ cDCs versus cDC1s (log2FC > 0.2 and adjusted *P*-value < 0.05) and cDC2-derived *LAMP3*$^+$ cDCs versus cDC2s (log2FC > 0.15 and adjusted *P*-value < 0.05). Then, we compared the two lists of DEGs and generated three groups of maturation associated genes: cDC1-specific, cDC2-specific and shared maturation signatures. Finally, the top 100 genes ordered by log2FC from each maturation signature were used as gene sets of interest for NicheNet analysis. The list of these genes was provided in Table S4A. All expressed genes in cDC1- or cDC2-derived *LAMP3*$^+$ cDCs were used as background of genes.

### Cell-cell interaction analysis

We used CellPhoneDB (Efremova et al., 2020) to infer cell-cell interaction between myeloid cell subsets. The potential interaction strength between two cell subsets was predicted based on the expression of ligand-receptor pairs. The enriched ligand-receptor interactions between two cell subsets were calculate based on permutation test. We extracted significant ligand-receptor pairs with *P*-value < 0.01.

### SCENIC analysis

Activated regulons in each cDC2 subset were analyzed using SCENIC (Aibar et al., 2017) with raw count matrix as input. Briefly, the co-expression network was calculated by *GRNBoost2* and the regulons were identified by *RcisTarget*. Next, the regulon activity for each cell was scored by *AUCell*. The differentially activated regulons in each cDC2 subset were identified by *wilcox.test* using cells from other cDC2 subsets as control. The Benjamini-Hochberg procedure was used to correct multiple hypothesis.

### Cell identity predicted by SciBet

To extend the established landscapes of tumor-infiltrating myeloid cells to other scRNA-seq datasets, we used SciBet (Li et al., 2020), a supervised cell type identifier based on E-test, to predict cell identities for cells from other datasets, such as treatment-related datasets from melanoma patients. We also used the established cDC2 subsets from 8 cancer types to predict the identities of cDC2 cells from the remaining cancer types. The marker genes for each subset were selected by *SelectGene* function in SciBet.

### Evaluation of mixability with entropy metric

We used an entropy-based metric to quantify the mixability of myeloid sub-populations across different cancer type (Azizi et al., 2018). We randomly sampled 50,000 myeloid cells and constructed a k-NN graph (k = 80) using the R function *kNN* from *dbscan* package (version 1.1.2), based on the Euclidean distance of cells in UMAP coordinates generated by Scanorama, and defined the mixability of the data as Shannon entropy,

$$H_j = -\sum_{t=1}^{T} p_j^t \log_2 p_j^t$$

where $p_j^t$ is the ratio of the number of cells from cluster $t$ in the 80 nearest neighborhood of cell $j$ and $\sum_{t=1}^{T} p_j^t = 1$.

### Gene set variation analysis

The GSVA package from Bioconductor was used for gene set variation analysis. The gene set we investigated (*c2BroadSets*) was loaded from the GSEABase package. The pathways with high difference in activity scores were selected by LIMMA package.

### Bulk sequencing and data preprocessing

#### *Bulk DNA and RNA isolation and sequencing*

For patients collected by our lab, the genomic DNA of tumor tissue and peripheral blood were extracted using the QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer's specification. The concentrations of DNA were quantified using the Qubit HsDNA Kits (Invitrogen). Exon libraries were constructed using the SureSelectXT Human All Exon V6 capture library (Agilent). Samples were sequenced on the Illumina Hiseq X Ten sequencer with 150-bp paired-end reads. The RNA of the tumor and adjacent normal tissues were extracted using the RNeasy Mini Kit (QIAGEN) according to the manufacturer's specification. The concentrations of RNA were quantified using the NanoDrop instrument (Thermo) and the qualities of RNA were evaluated with fragment analyzer (AATI). Libraries were constructed using NEBNext Poly (A) mRNA Magnetic Isolation Module kit (NEB) and NEBNext Ultra RNA Library Prep Kit for Illumina Paired-end Multiplexed Sequencing Library (NEB). Finally, RNA samples were sequenced on the Illumina Hiseq X Ten sequencer with 150-bp paired-end reads.

### Bulk exome sequencing data analysis

We used the BWA-PICARD/GATK-strelka pipeline to call somatic mutations. First, the cleaned read pairs were aligned to human genome reference (version b37 with decoy sequence added, download from Broad) by the BWA-MEM algorithm (Li and Durbin, 2010). Then, the aligned reads were sorted and de-duplicated by Picard (http://broadinstitute.github.io/picard). GATK 3.8 (DePristo et al., 2011) was used to realign multiple reads around putative INDEL and re-calibrate base quality. Finally, the bam files were input into strelka (Saunders et al., 2012) to call somatic mutations, including both somatic SNVs and INDELs. ANNOVAR (Wang et al., 2010) was used to annotate the somatic mutations. We calculated Tumor mutation burden (TMB) for each patient based on dividing the number of somatic nonsynonymous mutations by the total length of sufficiently sequenced regions (the captured regions by exome assay with > 30 sequencing depth in both the tumor and matched normal sample) measured in megabase.

### Bulk RNA-seq analysis

We applied the UCSC Xena Toil RNA-seq pipeline (Vivian et al., 2017) to quantify gene expression level, with hg38 as the reference genome and gencode V23 as the genomic annotation. RSEM (Li and Dewey, 2011) was used to quantitate gene expression.

### Correlation between myeloid subset composition and other factors

We first interrogated the relationship between the proportion of myeloid subset and somatic mutation at both gene and pathway levels. Specifically, we performed lasso regression to select somatic mutations correlated with myeloid cell subset proportions at the gene level. Tumor mutation burden (TMB) and genes collected in Cancer Gene Census (CGC) (Sondka et al., 2018) with mutation rates larger than or equal 0.1 were used as factors for lasso regression by using *glmmLasso* function from glmmLasso R package.

Cancer type was used as a random effect and the parameter lambda was selected by using a 3-fold cross validation. At the pathway level, we aggregated somatic mutation profiles into hallmark gene sets downloaded from MSigDB (Liberzon et al., 2015), and then conducted lasso regression as described for the gene level analysis. Only genes or pathways with *P value* less than 0.05 for the lasso regression were visualized in the heatmap.

We also examined the correlation between proportion of myeloid subset and bulk RNA-seq gene expression level in tumor tissues. We performed this analysis in 5 cancer types (ESCA, PAAD, KIDNEY, THCA and UCEC) with sufficient samples. Using matched bulk RNA-seq data of tumor tissues, we calculated the Pearson correlation of myeloid subset fraction with gene expression in each cancer type (correlation analysis). The *P*-values were adjusted by Benjamini-Hochberg method. Then, for genes with adjusted *P*-value < 0.05 in the correlation analysis step, we further compared their expression levels across tumor and normal tissues by using Wilcoxon test to identified genes significantly and highly expressed in tumor tissues (differently expression (DE) analysis). Log2 fold change (log2FC) for each gene was calculated by subtracting the log2 transformed mean expression level in tumor and normal tissues. Finally, we filtered genes with log2FC large than 0.5 and *P*-value less than 0.05 in DE analysis.

### TCGA data analysis

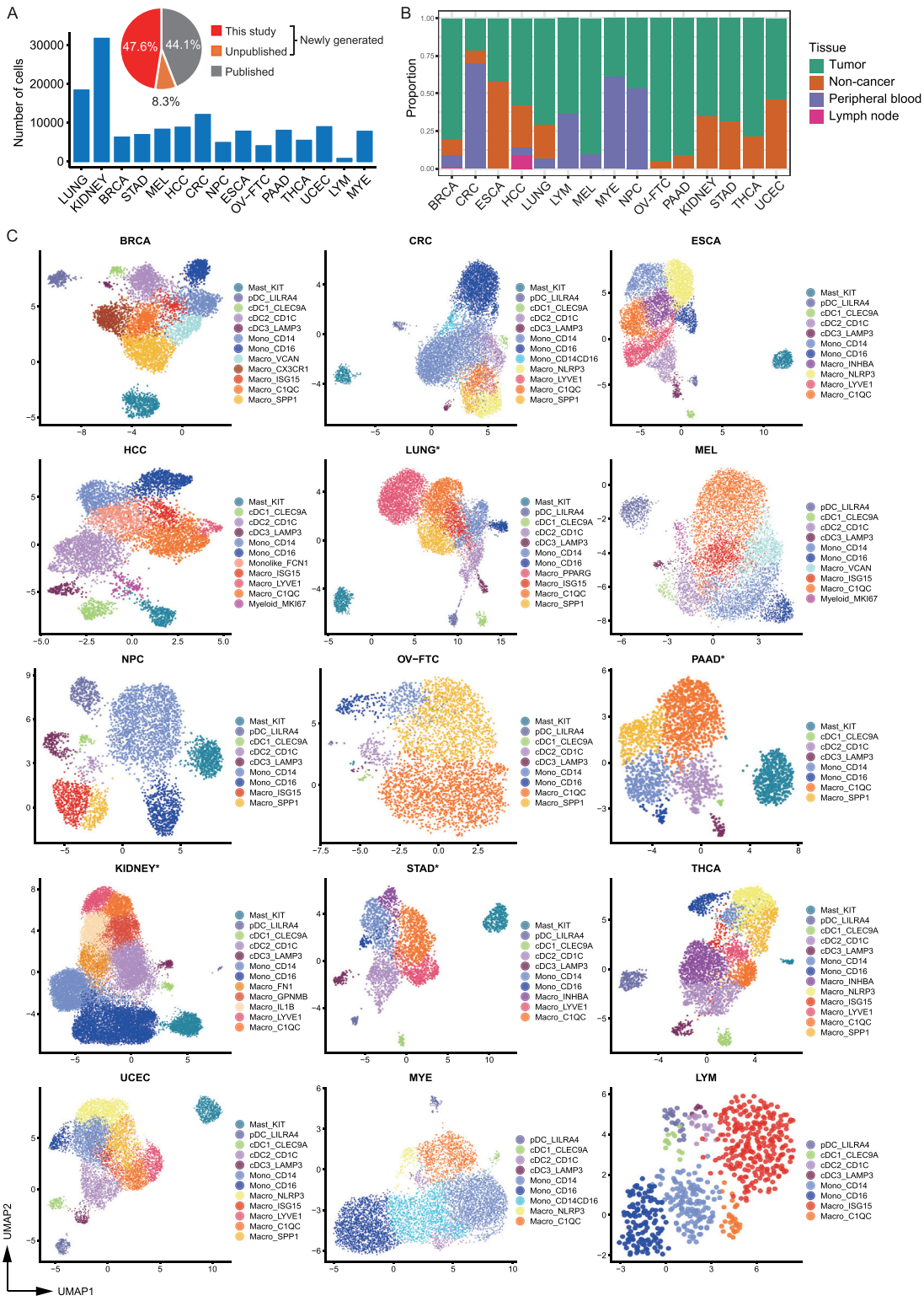
The expression data of TCGA were downloaded by using TCGAbiolinks package. Two endpoints (overall survival (OS) and disease-free interval (DFI)) from the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) (Liu et al., 2018) were used to analyze patients' clinical outcomes. We used the Cox proportional hazards model implemented in the survival package to test the correlation of selected genes and patents' survival. Kaplan-Meier survival curves were plotted by using R function *ggsurvplot*.

Specifically, to test the correlation of mast cells and patients' survival (using DFI), we used the four marker genes (*KIT*, *TPSAB1*, *CPA3* and *TPSB2*). To correct the effect of *CD45*$^{+}$ cell levels within each sample, the expression of selected genes in tumor were divided by the expression of *PTPRC* (*CD45*). The mean expression of the four genes was used to group samples into high and low groups based on 60$^{th}$ and 40$^{th}$ percentile respectively. Cox model was used to correct age, sex and tumor stage for survival analysis. Notably, for OV and UCEC, we only corrected age for survival analysis.

To test the association between angiogenesis-associated macrophage and patients' clinical outcomes (using OS), we used a single maker gene. For *SPP1*, we corrected the expression by *PTPRC* and then grouped samples into high and low groups based on 60$^{th}$ and 40$^{th}$ percentile respectively. For *FN1* and *VCAN*, we used the real expression and divided samples into two groups. Then we used the Cox proportional hazards model to correct age, sex and tumor stage for survival analysis.

CellPress
OPEN ACCESS

# Supplemental Figures



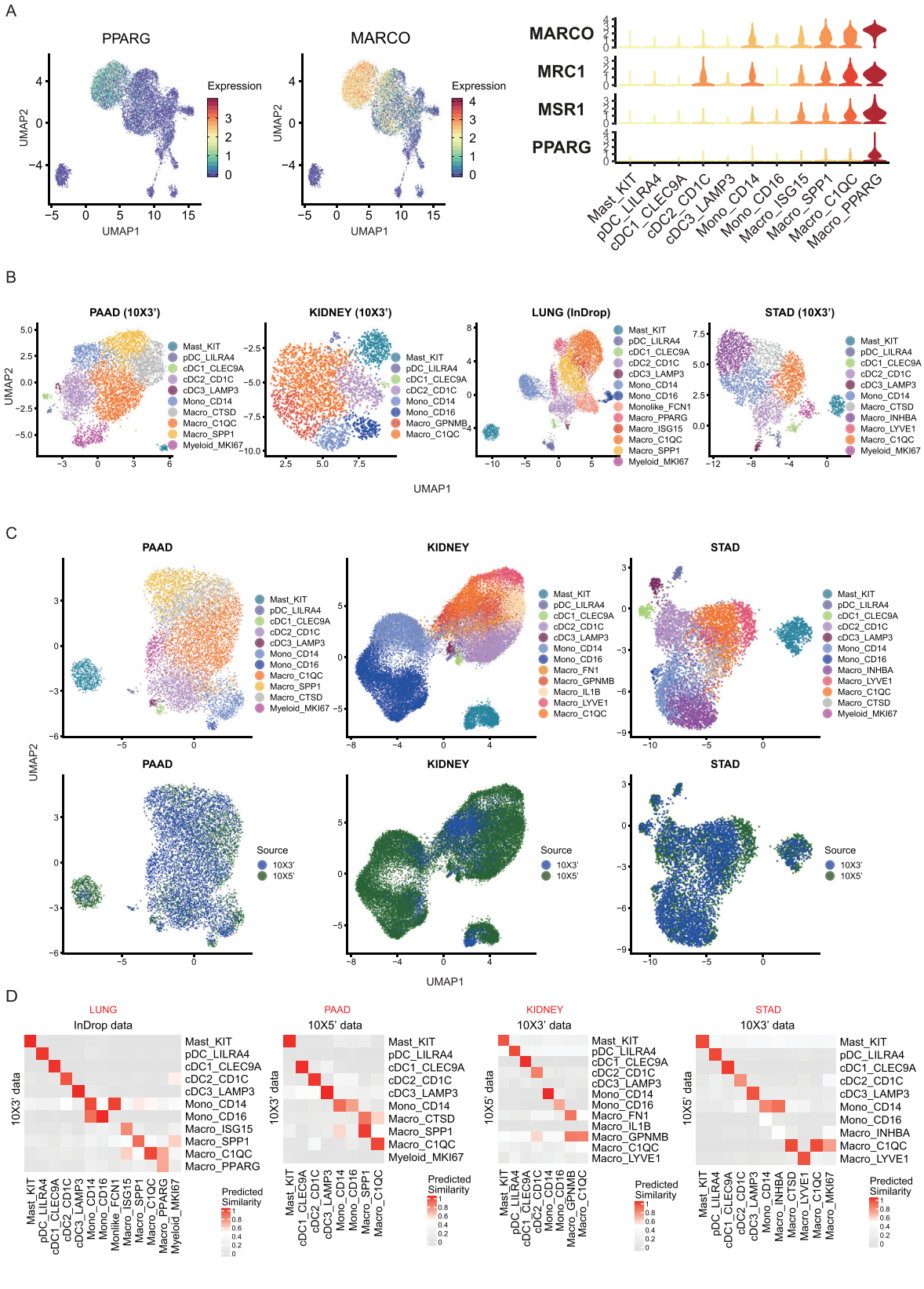

*(legend on next page)*

**Figure S1. Basic information of the data and myeloid cell subsets in each cancer type, related to Figure 1**

(A) Bar plot showing the number of myeloid cells collected in each cancer type. Pie chart showing the proportion of myeloid cells from different data sources.

(B) The fraction of cells originating from each of the 4 tissues across multiple cancer types.

(C) UMAP plots showing the subsets of myeloid cells in the 15 cancer types. *, cancer types comprising two independent datasets.




*(legend on next page)*

**Figure S2. Analysis of myeloid cell subsets for four cancer types with datasets from two independent studies, related to Figure 1**

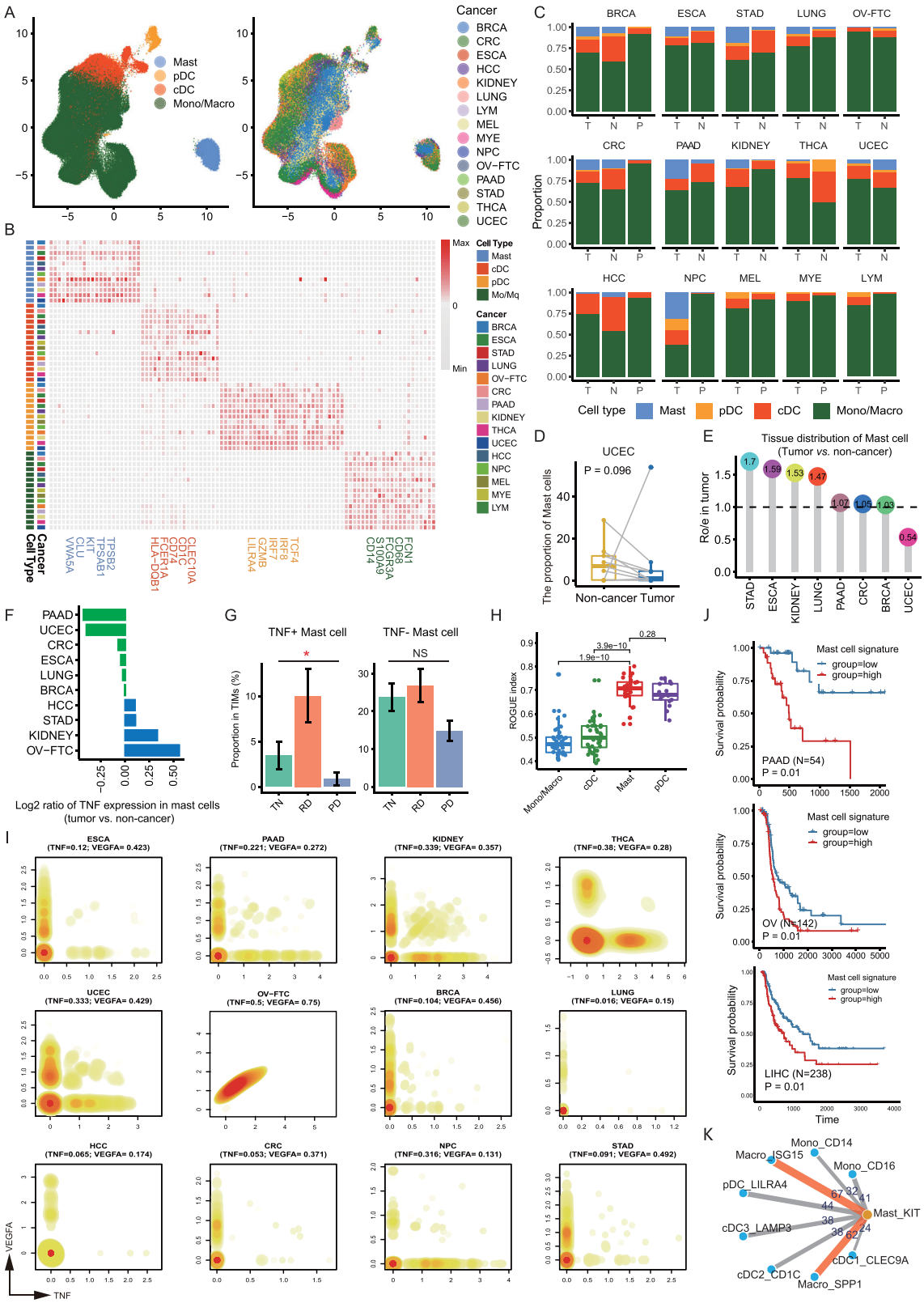
(A) UMAP and violin plots showing the expression of *PPARG*, *MARCO*, *MRC1* and *MSR1* in lung cancer macrophage subsets.

(B) UMAP plots showing the subsets of myeloid cells in an independent dataset for the four cancer types (LUNG, PAAD, KIDNEY and STAD).

(C) UMAP plots showing Scanorama's integration of two independent datasets.

(D) Heatmaps showing similarity of myeloid cell subsets identified from two datasets.

CellPress
OPEN ACCESS




*(legend on next page)*

**Figure S3. Characteristics of major myeloid cell lineages in various cancer types, related to Figure 2**

(A) UMAP plots showing the combination of all datasets by Scanorama.

(B) Gene expression heatmap showing the common signature genes for major cell lineages.

(C) The distribution of the major cell lineages in each tissue. N, non-cancer; T, tumor; P, peripheral blood.

(D) Boxplot showing the comparison of mast cells fraction (divided by the total myeloid cell number) in tumor and non-cancer tissues from UCEC patients. Paired two-sided Wilcoxon test.

(E) Lollipop plot showing the tissue distribution of mast cells by *Ro/e* analysis in various cancer types (with more than 200 mast cells).

(F) Bar plot showing the ratio of *TNF* expression level in mast cells originated from tumor to non-cancer tissues.

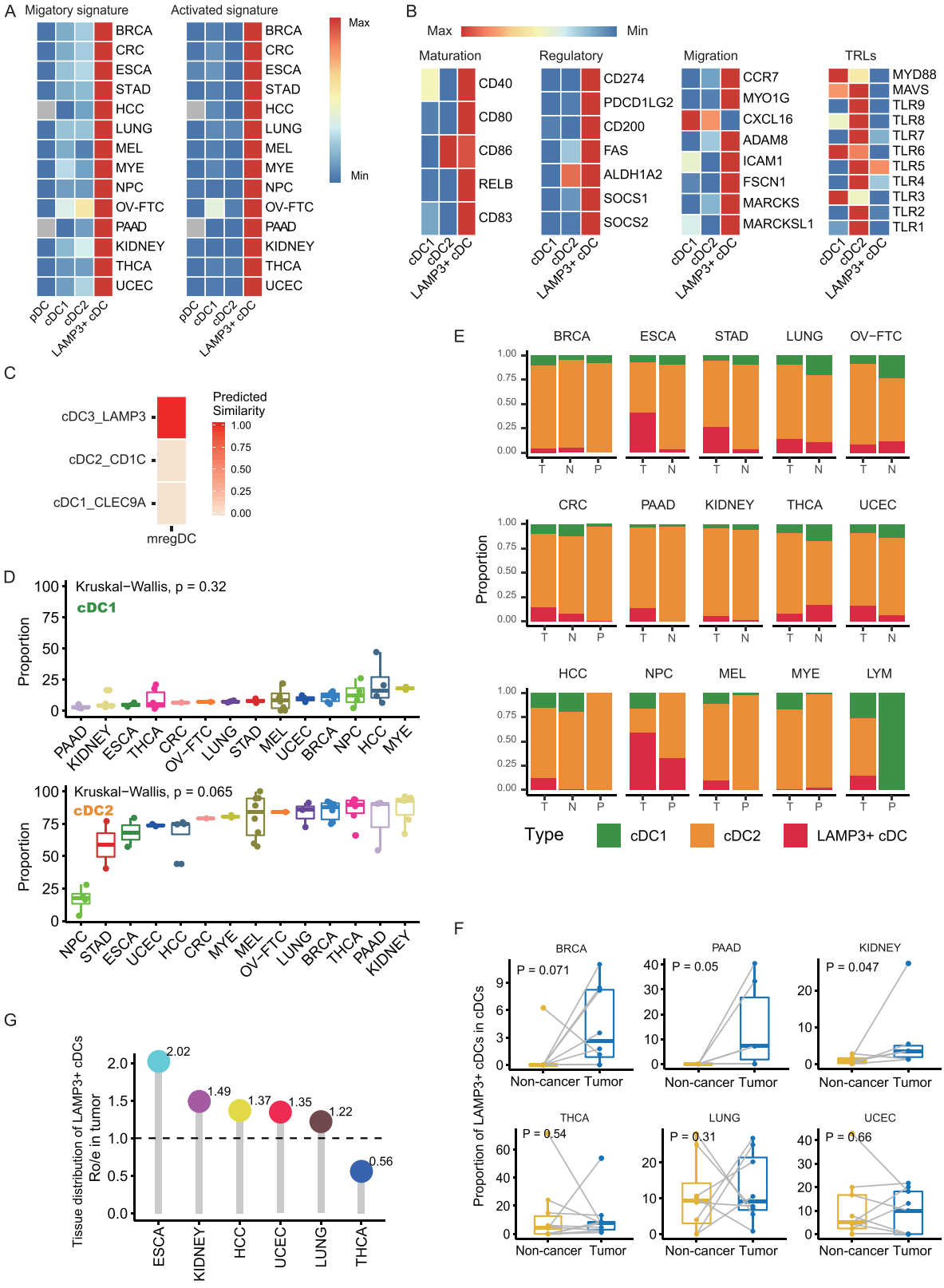
(G) Bar plot showing fractional changes for $TNF^-$ and $TNF^+$ mast cells in TIMs across the three treatment states. The proportion was calculated by dividing the total cell number of myeloid cells. Error bars indicate the 95% confidence interval for the calculated relative frequencies. *p < 0.01 using a chi-square test of independence.

(H) Boxplot showing cell purity for each major myeloid lineage by ROGUE. Unpaired two-sided Wilcoxon test.

(I) Expression patterns of *TNF* and *VEGFA* in mast cells across different cancer types.

(J) Kaplan-Meier plots showing worse clinical outcome in pancreatic adenocarcinoma (PAAD), ovarian serous cystadenocarcinoma (OV) and liver hepatocellular carcinoma (LIHC) patients with the higher expression of mast cell marker genes. +, censored observations. *P*-value was calculated by multivariate Cox regression.

(K) Network showing number of significant interaction events between mast cells with other myeloid cell subsets in NPC.




*(legend on next page)*

**Figure S4. Characteristics and tissue distribution pattern of *LAMP3*$^+$ cDCs, related to Figure 3**

(A) Heatmaps showing the migratory and activated signatures in different DC subsets.

(B) Heatmaps showing the signatures of 'mregDC' in different DC subsets.

(C) Heatmap showing similarity between 'mregDC' and cDC subsets.

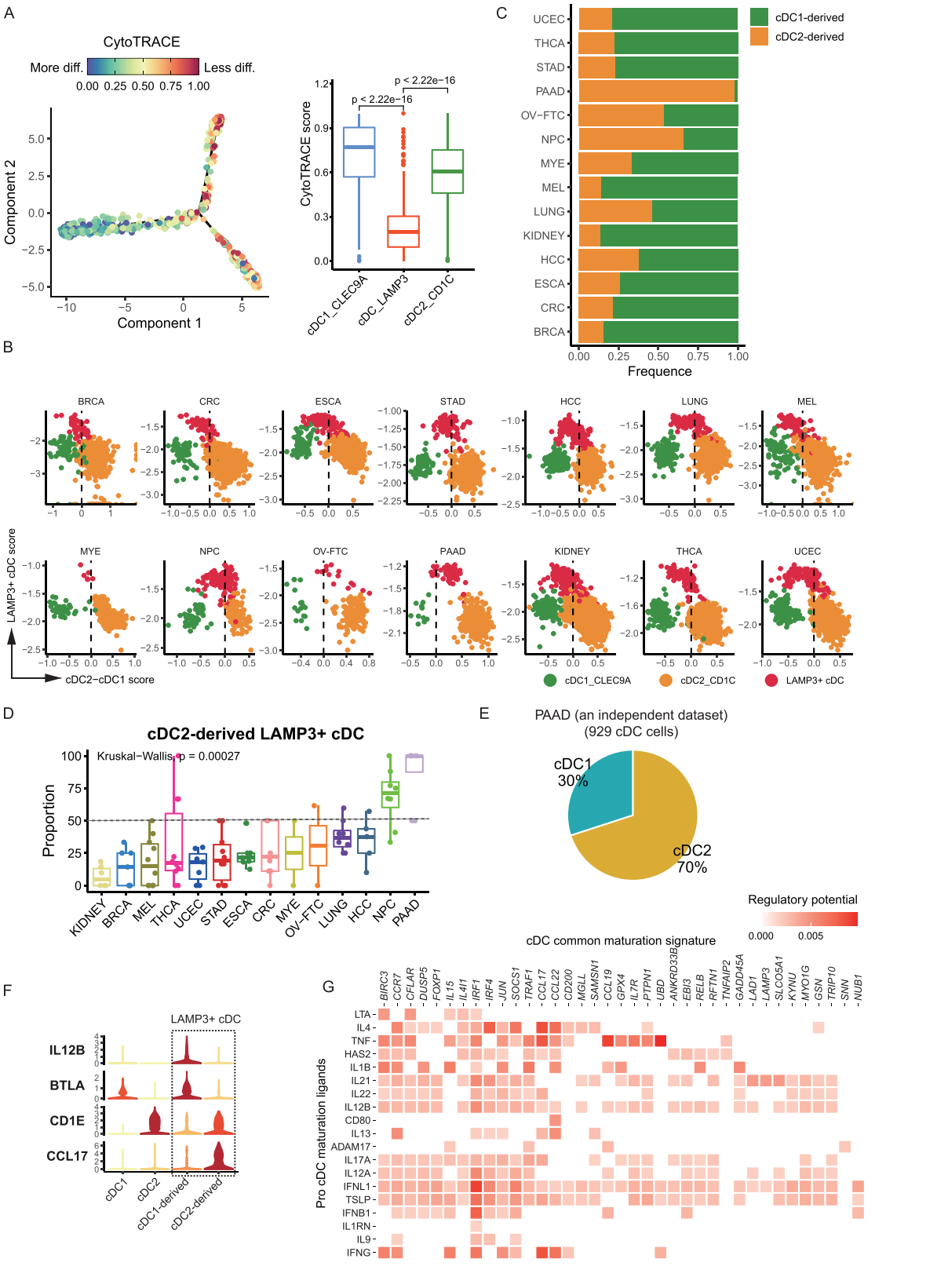
(D) Boxplots showing the proportions of cDC1 and cDC2 subset (divided by the total cell number of cDCs) in tumor tissues across various cancer types. Kruskal-Wallis test.

(E) The tissue distribution of *LAMP3*$^+$ cDCs in different tissues. N, non-cancer; T, tumor; P, peripheral blood.

(F) Boxplots showing the comparison of *LAMP3*$^+$ cDCs fraction in tumor and non-cancer tissues. Paired two-sided Wilcoxon test.

(G) Lollipop plot showing the tissue distribution of *LAMP3*$^+$ cDCs by *Ro/e* score.

We only included cancer types with more than 5 *LAMP3*$^+$ cDCs in (A) and (D).




*(legend on next page)*

**Figure S5. Origins and regulation of *LAMP3*$^+$ cDCs, related to Figure 4**

(A) CytoTRACE analysis of the origins of *LAMP3*$^+$ cDCs. Combined application of CytoTRACE and Monocle2 to predict the origins of *LAMP3*$^+$ cDCs (Left). Boxplot showing the comparison of CytoTRACE score between different cDC subsets (Right). Unpaired two-sided Wilcoxon test.

(B) Stratification of cDC transcriptomes by scores generated from signature genes of three cDC subsets.

(C) Bar plot showing the origins of *LAMP3*$^+$ cDC in each cancer predicted by SingleR.

(D) Boxplot showing the fractions of cDC2-derived *LAMP3*$^+$ cDCs in all *LAMP3*$^+$ cDCs across different cancer types. Kruskal-Wallis test.

(E) Pie chart showing the origins of *LAMP3*$^+$ cDC in PAAD using an independent dataset.

(F) Violin plot showing the expression of *IL12B*, *BTLA*, *CD1E* and *CCL17* in cDC1- and cDC2- derived *LAMP3*$^+$ cDCs.

(G) Heatmap showing shared potential ligands driving the maturation of both cDC1s and cDC2s.

We only included cancer types with more than 5 *LAMP3*$^+$ cDCs in (B)–(D).

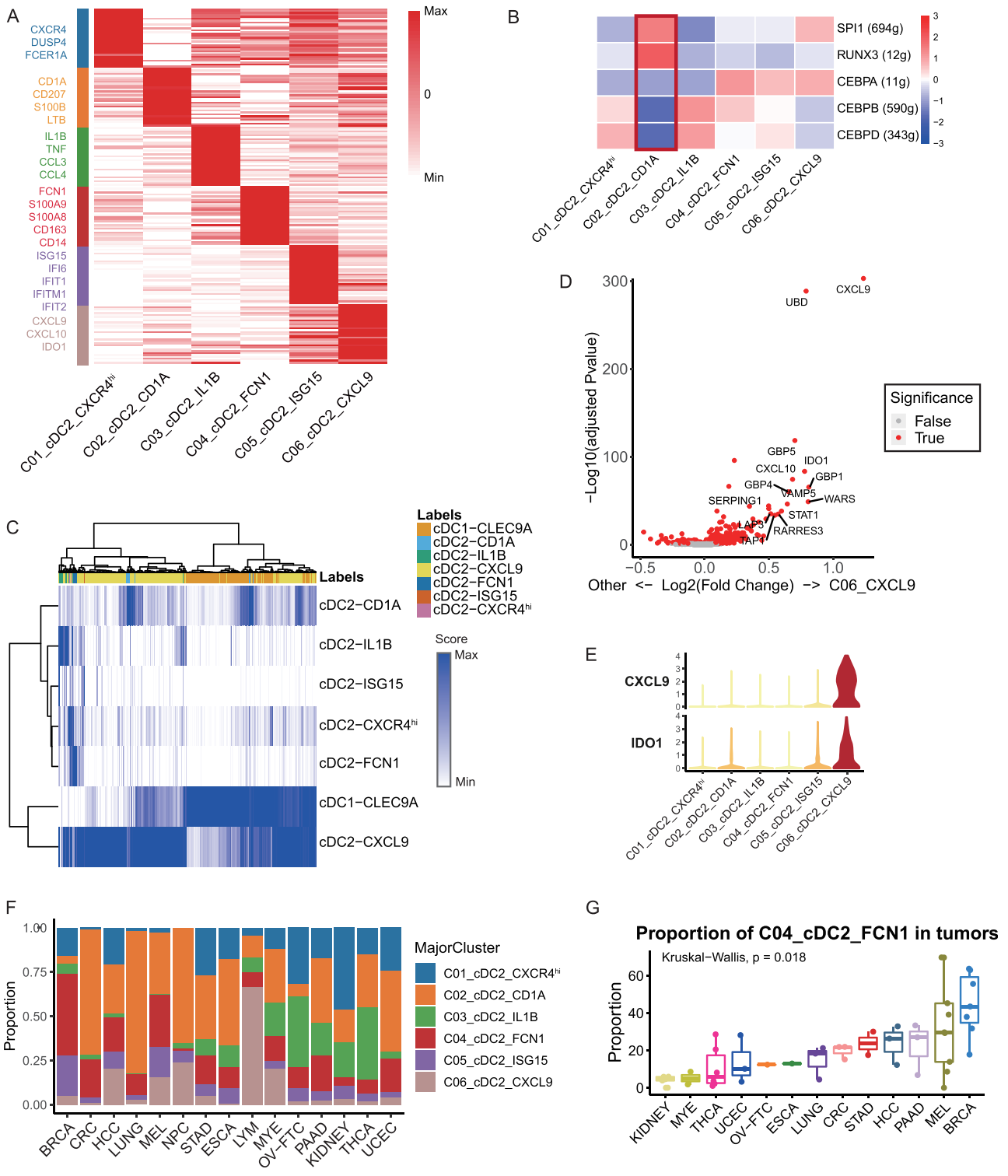



**Figure S6. Characteristics of cDC2 subsets, related to Figure 5**

(A) Heatmap showing the signature genes for each cDC2 subset.

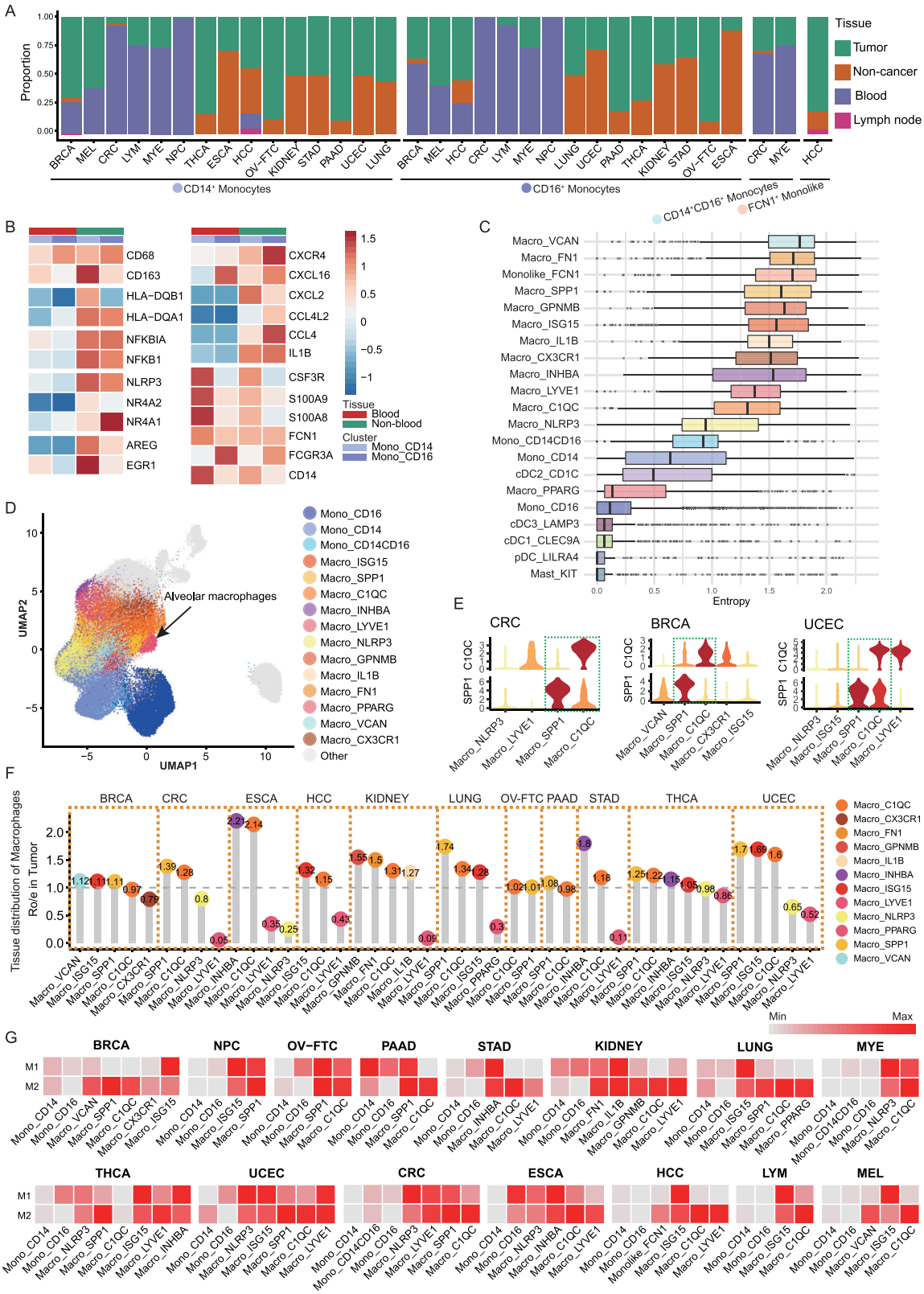
(B) Heatmap showing the activity of Langerhans cell-associated TFs across 6 cDC2 subsets.

(C) Heatmap showing the origins of *LAMP3*$^{+}$ cDCs predicted by SingleR. Score displayed in the plot represents probability for cell assigned into a specific cluster.

(D) Volcano plot showing the differentially expressed genes between C06_cDC2_CXCL9 sub-cluster and other cDC2 sub-clusters. Adjusted *P*-value < 0.05, Benjamini-Hochberg adjusted two-sided Wilcoxon test.

*(legend continued on next page)*

(E) Violin plot showing the expression of *CXCL9* and *IDO1* across 6 cDC2 subsets.
(F) Bar plot showing the fraction of each cDC2 subsets in tumor tissues across various cancer types.
(G) Boxplot showing the proportions of C04_cDC2_FCN1 (divided by the total cell number of cDC2s) in tumor tissues across various cancer types. Kruskal-Wallis test.




*(legend on next page)*

**Figure S7. Characteristics of monocytes and integration of monocyte and macrophage cells in various cancer types, related to Figure 6**

(A) The fraction of cells originating from each of the 4 tissues across all myeloid cell sub-populations.

(B) Heatmaps showing genes expressed differently between blood derived monocytes and non-blood derived monocytes. This plot was generated with corrected expression value from Scanorama.

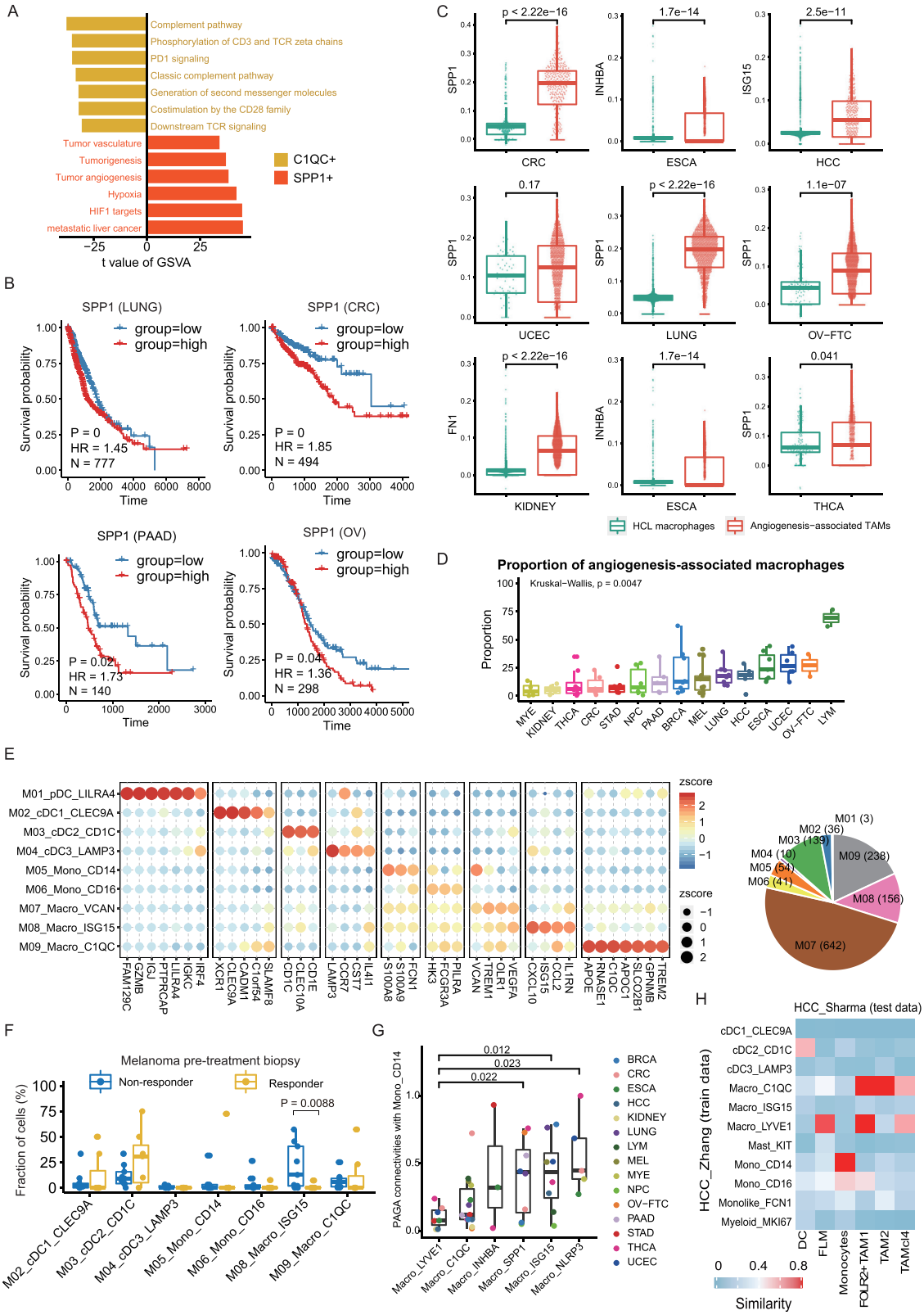
(C) Boxplot showing the mixability of each myeloid subset evaluated by an entropy-based metric.

(D) UMAP plot showing the integration of monocyte and macrophage sub-clusters by Scanorama.

(E) Violin plots showing the distinct expression patterns of *SPP1* and *C1QC* in different cancer types.

(F) Lollipop plot showing the tissue distribution of each macrophage subset in tumor by *Ro/e* analysis. Tumor-enriched subsets were characterized with *Ro/e* > 1.

(G) Heatmaps showing the M1 and M2 signatures across all monocyte and macrophage subsets in each cancer type.




*(legend on next page)*

**Figure S8. Characteristics of macrophage subsets in various cancer types, related to Figure 6**

(A) Bar plot showing different pathways enriched in *C1QC*$^{+}$ macrophage and *SPP1*$^{+}$ macrophage in lung cancer scored per cell by gene set variation analysis (GSVA). t values are calculated with limma regression.

(B) Kaplan-Meier plots showing worse clinical outcome in four types of cancer patients (LUNG, PAAD, OV and CRC) with the higher expression of *SPP1*. +, censored observations; HR, hazard ratio. *P*-value was calculated by multivariate Cox regression.

(C) Boxplot showing the expression levels of marker genes of angiogenesis-associated TAMs and macrophages in corresponding normal tissues from HCL. Unpaired two-sided Wilcoxon test.

(D) Boxplot showing the proportions of angiogenesis-associated macrophages (labeled with "*" in Figure 6E, divided by the total myeloid cell number) across various cancer types. Kruskal-Wallis test.

(E) Bubble heatmap showing expression levels of selected marker genes for each myeloid subset in treatment-related dataset (left). Pie chart showing cell numbers of each myeloid cell subset in treatment-related dataset (right).

(F) Boxplot showing the proportion of each myeloid cell subset divided by the total myeloid cell number in patients with different immunotherapy responses to checkpoint inhibitors immunotherapy. Unpaired one-sided Wilcoxon test.

(G) Boxplot showing the connectivity between different macrophage clusters and CD14+ monocytes. Unpaired two-sided t test.

(H) Heatmap showing the similarities across different clusters in two HCC datasets.