HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2020 August 10.

Published in final edited form as:

Nat Med. 2020 February ; 26(2): 259–269. doi:10.1038/s41591-019-0750-6.

Regenerative lineages and immune-mediated pruning in lung cancer metastasis

Ashley M. Laughney[1,2,3,4,5], **Jing Hu**[1], **Nathaniel R. Campbell**[1,2,6], **Samuel F. Bakhoum**[7,8], **Manu Setty**[2], **Vincent-Philippe Lavallée**[2], **Yubin Xie**[2,9], **Ignas Masilionis**[2,10], **Ambrose J. Carr**[2], **Sanjay Kottapalli**[2,10], **Viola Allaj**[11,12], **Marissa Mattar**[11,12], **Natasha Rekhtman**[13], **Joao B. Xavier**[2], **Linas Mazutis**[2,10], **John T. Poirier**[14], **Charles M. Rudin**[11,12], **Dana Pe'er**[2,15,*], **Joan Massagué**[1,*]

[1]Cancer Biology and Genetics Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[2]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[3]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, New York, USA

[4]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, USA

[5]Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, New York, USA

[6]Tri-Institutional MD-PhD Program, Weill Cornell/Rockefeller University/Sloan Kettering Institute, New York, New York, USA

[7]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[8]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[9]Tri-Institutional Training Program in Computational Biology and Medicine, Weill Cornell/ Rockefeller University/Sloan Kettering Institute, New York, New York, USA

[*]co-communicating co-senior authors: j-massague@ski.mskcc.org and peerd@mskcc.org.

Author Contributions. A.M.L., J.M. and D.P. conceived and oversaw the project, performed data interpretation and wrote the manuscript. A.M.L. generated and analyzed all scRNA-seq data and performed mouse experiments for single cell analyses; J.H. generated additional NK depleted macrometastases for immunofluorescence, which were processed by A.M.L. and J.H. J.H. and S.F.B. identified matched primary-metastasis pairs, analyzed patient immunofluorescence, and assisted with data interpretation. J.H. designed and performed NK cell cytotoxicity assays *in vitro*. N.R.C. and J.B.X. developed and applied image cytometry methods for NK cell cytotoxicity assays *in vitro*. M.S., V.-P.L., A.J.C. and S.K. assisted with scRNA-seq analysis. Y.X. developed and applied algorithms to segment nuclei in fixed tissues. N.R., V.A., M.M., J.T.P. and C.M.R. oversaw procurement of tissue samples, clinical specimen processing and histopathologic data interpretation. I.M. and L.M. assisted with scRNA-seq library preparation.

Competing Interests

J.M. is a scientific advisor and owns company stock in Scholar Rock. C.M.R. has consulted with AbbVie, Amgen, Ascentage, Astra Zeneca, BMS, Celgene, Daiichi Sankyo, Genentech/Roche, Ipsen, Loxo, and Pharmar, and is on the scientific advisory boards of Elucida and Harpoon. S.F.B. owns equity in, receives compensation from, and serves as a consultant, board member, and a scientific advisory board member for Volastra Therapeutics Inc. He also has consulted for Sanofi. All other authors declare no competing conflicts.

[10]The Alan and Sandra Gerry Metastasis and Tumor Ecosystems Center, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[11]Molecular Pharmacology Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[12]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[13]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[14]Perlmutter Cancer Center, New York University Langone Health, New York, New York, USA

[15]Parker Institute for Cancer Immunotherapy, Memorial Sloan Kettering Cancer Center, New York, New York, USA

Abstract

Developmental processes underlying normal tissue regeneration have been implicated in cancer, but the degree of their enactment during tumor progression and under the selective pressures of immune surveillance, remain unknown. Here, we show that human primary lung adenocarcinomas are characterized by the emergence of regenerative cell types typically seen in response to lung injury, and by striking infidelity amongst transcription factors specifying most alveolar and bronchial epithelial lineages. In contrast, metastases are enriched for key endoderm and lung-specifying transcription factors, *SOX2* and *SOX9*, and recapitulate more primitive transcriptional programs spanning stem-like to regenerative pulmonary epithelial progenitor states. This developmental continuum mirrors the progressive stages of spontaneous outbreak from metastatic dormancy in a mouse model and exhibits *SOX9*-dependent resistance to Natural Killer (NK) cells. Loss of developmental stage-specific constraint in macrometastases triggered by NK cell depletion suggests a dynamic interplay between developmental plasticity and immune-mediated pruning during metastasis.

Tissue homeostasis is maintained by stem cells[1], whereas damaged tissues are repaired by facultative progenitors that are activated upon injury[2–5]. In the pulmonary epithelium, a tissue with relatively low turnover, most epithelial cell types can interconvert in the context of growth or injury[6,7]. There is evidence that such plasticity extends to primary lung cancer, in which embryonic pathways that typically control stem and progenitor cell behavior become active[8,9]. Yet, the role of developmental plasticity in tumor progression and metastasis remains poorly understood[10]. In particular, the extent to which tumor cells subvert regenerative processes during metastatic progression, and how this is shaped by the sustained selective pressures of anti-tumor immunity, is unknown. We used droplet-based single-cell RNA sequencing (scRNA-seq)[11–13] and graph-based phenotypic analysis[14,15] to explore tumor cell heterogeneity through the lens of epithelial regeneration in lung cancer metastasis[16], and to assess parallels between tumor cell plasticity and developmental hierarchies[17]. Using patient tumors as well as a mouse model of lung cancer metastasis, our analyses identify regenerative cell types and lineage promiscuity in untreated primary tumors and reveal a range of embryonic lung morphogenic states in metastases. We demonstrate an unexpected, developmental stage-specific differential sensitivity to NK cells that shapes the phenotypic landscape of overt colonization from latent metastasis-initiating

cells. Depletion of NK cells allows diverse cancer cell types that were targeted to emerge during metastasis. In this context, *SOX9* expression in cancer cells confers resistance to NK cell-mediated killing, suggesting an intimate relationship between lineage-determining transcriptional programs and immune vulnerabilities during lung cancer evolution.

Single-cell transcriptional landscape of primary and metastatic lung adenocarcinomas

We profiled the transcriptomes of 40,505 individual cells obtained from 17 freshly resected human tissue samples comprising adjacent non-tumor involved lung (n = 4, hereafter 'normal lung'), primary lung adenocarcinoma (LUAD, 7 untreated and 1 post neo-adjuvant chemotherapy), as well as three LUAD metastases from brain, one from bone, and one from adrenal glands (Fig. 1a). These samples spanned various stages of tumor progression (Extended Data Fig. 1). All the scRNA-seq data were merged and normalized to create a global cell atlas. Clustering[15] revealed 20 cell types spanning stromal, lymphoid, myeloid, epithelial and endothelial cells, pericytes and fibroblasts, as well as cancer cells (Fig. 1b–c). The library size, complexity and viability metrics were of high quality (Extended Data Fig. 2a–c) and largely consistent across patients (Extended Data Fig. 2d–e). Although their abundances varied by sample (Fig. 1b), most major myeloid, lymphoid and stromal cell types (annotated in Extended Data Fig. 3–4) were highly reproducible across patients (Extended Data Fig. 5a–b), and were detectable in both the merged global atlas and in individual patient samples (Extended Data Fig. 5d–e). In contrast, patient-specific cell states emerged within cancer cells of the neo-adjuvant-treated primary tumor and metastases, suggesting biological selection in later stage disease.

Cell type assignments were further refined within myeloid, epithelial, and stromal compartments (Extended Data Fig. 3) separately from the lymphoid compartment (Extended Data Fig. 4), to avoid biases introduced by cell-type-specific capture rates. The frequency of these cell types varied significantly (Kruskal-Wallis rank test) between normal lung, primary tumors, and metastases (Fig. 1d). For example, NK cells were depleted in primary LUADs compared to normal lung (Fig. 1d), consistent with a recent report[18]. Myeloid cells were distinguished by upregulation of inflammation, wound healing and antigen presentation genes, whereas stromal cell types were highly proliferative and expressed BMP, FGF and WNT cytokines implicated in lung morphogenesis (Extended Data Fig. 5a–b). Epithelial and carcinoma subpopulations exhibited distinct expression patterns of transcription factors and surface markers characteristic of specialized lung epithelial cell types including an alveolar epithelial progenitor (AEP) known to regenerate the alveolar epithelium upon injury[5] (Extended Data Fig. 3b and Extended Data Fig. 7). Of note, the majority of sequenced cells were immune, with cancer cell purity ranging from 7–32% per sample. These estimates were slightly lower than purity estimates[19] of 20–40% per sample from bulk targeted gene sequencing[20] (Extended Data Fig. 5c, $R = 0.58$, $p = 0.08$), likely due to a bias in epithelial cell recovery. However, even bulk estimates point to a similarly high level of immune infiltrate in our cohort and in other studies of non-small cell lung cancer[21].

Regenerative and mixed lineages in primary tumors

We next focused on the relationship between primary carcinoma cells and adult lung epithelial lineages, which are directed to specialize by the expression of transcription factors *SOX2* and *SOX9* during development[22] (Fig. 2a). Within the adult normal epithelium, we identified four mature cell types including alveolar epithelial cell types 1 and 2 (AEC1 and AEC2) of the distal lung, as well as ciliated and club cells of the upper airway (Extended Data Fig. 6a–b). Lineage-specific gene sets defined in the developing mouse embryo at D18.5 [23–25] (Supplementary Table 1) distinguished these four epithelial cell types, which showed near-mutual exclusivity in their top differentially expressed genes (DEGs) (Extended Data Fig. 6c–d). Canonical markers of the lung epithelium were among the top DEGs (Extended Data Fig. 6c) and the majority of cells within the AEC1, AEC2 and ciliated clusters highly expressed more than 60% of genes within each lineage-specific gene signature (Extended Data Fig. 6e).

Analysis of the combined normal and primary tumor epithelium revealed the emergence of regenerative and mixed-lineage states in cancer (Fig. 2, Extended Data Fig. 7a–b). In addition to the four mature lineages identified in normal lung, two progenitor cell types implicated in the regeneration of severely injured lung were detected in primary tumors (Fig. 2a–c). These include *SOX2*-derived *KRT5*+ basal-like cells[3,4] (Phenograph Cluster 13), which exhibited increased RAS signaling and mesenchymal gene enrichment associated with wound response, and *SOX9*-expressing AEPs[5] (Phenograph Clusters 10–11), which were tumor-specific and predominantly derived (80%) from the neoadjuvant-treated primary tumor MSK-LX679 (Extended Data Fig. 7a–c). We also found that *SOX9*-expressing AEP-like cells are frequently observed in advanced, LUAD metastases (see below). Interestingly, cells annotated as AEPs also expressed genes associated with mucin production, supporting the notion that cells exhibiting progenitor activity may also retain differentiated programs[26] (Extended Data Fig. 7c). A third patient (MSK-LX680)-specific tumor cluster expressed canonical neuroendocrine (NE) markers (Extended Data Fig. 7c), consistent with this patient's diagnosis of LUAD mixed with large cell neuroendocrine carcinoma (Extended Data Fig. 1). Nearly all carcinoma-specific cell types showed increased expression of transcription factors associated with progenitor cell function (*ID2*, *SOX2* and *SOX9*, Extended Data Fig. 7c). The remaining six clusters, comprising more then 50% of all cancer cells, aberrantly expressed signatures related to multiple proximal and distal epithelial cell lineages and were therefore annotated as mixed-lineage (grey cells, Fig. 2b–d, Extended Data Fig. 7b). Top DEGs within these mixed-lineage clusters include markers for AEC1, AEC2, ciliated, AEP and basal-like cells (Fig. 2e, Extended Data Fig. 7a).

To further investigate lineage promiscuity in primary tumors at the level of individual cells, we compared the fraction of lineage-specific genes expressed in normal epithelial lineages and in mixed-lineage tumor cell clusters based on un-imputed data (Methods). The distribution of cells expressing pairwise combinations of lineage markers illustrates this promiscuity: densities of AEC1 and AEC2 normal cells overlap with mixed-lineage tumor cells (clusters 0 and 1) (Fig. 2f), demonstrating tumor cell promiscuity across these two alveolar cell types.

Quantifying transcription factor promiscuity across multiple epithelial lineages is a challenging combinatorial problem given the number of cell types and their dependencies. Therefore, we applied a phenotypic volume metric[13] (Methods) to capture complexity within the lineage-specific gene-gene covariance structure of tumor and normal cells. The expansion of lineage phenotypic volume in primary tumor cells compared to normal lung quantifies the greater heterogeneity in primary tumors due to aberrant lineage marker combinations (Fig. 2e–g).

This increased lineage plasticity in tumor cell states was nevertheless structured and bounded by predominantly normal, lung epithelial cell types (AEC1, AEC2 and club cells, Extended Data Fig. 7d–f). Specifically, the mixed-lineage clusters occupied intermediate states between boundary states identified as the extrema of the first three diffusion components[13,14,27,28]. While the extrema of diffusion components 1 and 3 were associated with the specification of mature AEC1, AEC2 and club cell types, diffusion component 2 more generally reflected proximal-distal patterning in the lung and showed enrichment of embryonic stem cell programs in proximal cell types (Extended Data Fig. 7d). The latter likely reflects the dual role of *SOX2* in lung endoderm specification and proximal patterning[29]. Importantly, tumor cells were similar to normal lung epithelium but also distinguished by the expression of an adenocarcinoma signature that is absent in non-cancerous lung epithelium[30] (Extended Data Fig. 7g), supporting their aberrant nature.

Our analysis thus shows that primary tumors exhibit lineage diversity and that tumor cell transcriptional profiles can resemble heterogeneous mixed lineage and regenerative cell types that expand in response to lung injury.

Developmental continuum in human metastases

Phenograph clustering of the combined normal, primary tumor and metastatic epithelium revealed 21 distinct phenotypic states based on genome-wide expression patterns, 11 of which were appreciably detected in metastases (Extended Data Fig. 8a–d). The metastatic clusters were distinguished by upregulation of pathways related to embryonic stem cells, wound healing and morphogenesis, as well as gene sets defining the conserved alveolar epithelial progenitor (Extended Data Fig. 8a). Therefore, we ranked the 11 metastatic clusters based on the average expression of a specific lung epithelial development signature (Fig. 3a, GO:0060428 listed in Supplementary Table 2) and explored their association with early lung specification and development programs (Fig. 3b). Based on the expression of multiple gene signatures, we highlight three metastatic states along this developmental progression (type I-III). The ordering of metastatic clusters by lung development negatively correlated with the expression of a recently described, adult stem cell signature linked to aggressive epithelial cancers[31] (Fig. 3a). Clusters with elevated adult stem cell signatures (type I and II) also exhibited higher expression of cell migration genes, and intermediate type II cells were enriched for morphogenesis and respiratory endoderm specification pathways. In contrast, cells from type III clusters exhibited the highest levels of AEP programs (Fig. 3a).

Next, we used the same lung epithelium signature to rank individual tumor cells. We found that the expression of key endoderm and lung-specifying transcription factors[32] likewise supports a phenotypic progression from foregut endoderm development (type I, *SOX17* and *HHEX*[33,34]), to lung and trachea morphogenesis (type II, *SOX2*, *NKX2–1* and *FOXA2*), to the WNT-responsive alveolar progenitors that give rise to all proximal and distal cell types (type III, *SOX9* and *WNT7B*) (Fig. 3c). To specifically explore developmental gene expression as an important source of variation, we used the single-cell imputed expression matrix of these genes exclusively (instead of the full transcriptome) to cluster tumor cells and identified four developmental stages: a proliferating (type I-P) and quiescent (type I-Q) adult stem-like state, a *SOX2*^{high} regenerative state (type II) and a *SOX9*^{high} alveolar epithelial progenitor state (type III).

Lung development is largely specified by *SOX2* and *SOX9*[29], and *SOX2* additionally directs embryonic development and foregut endoderm specification (Fig. 3b). Few cells in the late embryonic and postnatal stages retain expression of these transcription factors outside of an injury response[35,36]. We used immunofluorescence to analyze these critical factors in four matched primary-metastasis pairs, and observed strong nuclear SOX2 and SOX9 expression in metastatic epithelium that was undetectable in primary tumors, thereby corroborating single-cell transcriptional data showing increased expression of stem and regenerative programs in metastases (Fig. 3d and Extended Data Fig. 8c).

We asked whether these metastatic subpopulations can also be found in primary LUADs, and observed that cells corresponding to stem and regenerative (type I/II) metastatic stages exist to varying extents in nearly all primary tumors (Fig. 3e). Moreover, the abundance of type I cells was associated with reduced overall survival in LUAD patients[37] (Fig. 3f). This finding agrees with previous reports that stem cell transcriptional programs are associated with inferior outcomes across cancer types[31]. Conversely, *SOX9*-expressing AEP clusters showed additional patient-specific features, predominantly after neo-adjuvant chemotherapy (Extended Data Fig. 8a,e).

Collectively, these data suggest that metastases recapitulate key stages of endoderm and lung morphogenesis, which, when mapped to normal development (Fig. 3b), lie upstream of mature proximal and distal airway lineages that demarcate the phenotypic landscape of primary tumors.

Developmental continuum in mouse model of metastasis

While phenotypic states spanning key stages of endoderm and lung development are detected in patient metastasis, they are variably abundant and sampled at the endpoint of a dynamic, evolutionary process with multiple bottlenecks[10]. Therefore, we sought to recapitulate these findings in a transplantable mouse model of lung cancer metastasis derived from an early stage RAS-mutant human LUAD (H2087-LCC)[16], which may be interrogated over time. We previously showed that upon inoculation into the arterial circulation of athymic mice, H2087-LCC cells infiltrate multiple organs, persist as latent disseminated tumor cells (DTCs), express stem-like transcriptional signatures, including *SOX2* and *SOX9*, and retain metastasis-initiating potential[16]. They remain quiescent for

extended periods of time, infrequently leading to spontaneous macrometastatic outbreaks, which can be monitored using a bioluminescence (BLI) reporter. Treatment with antibodies that eliminate NK cells leads to extensive multi-organ metastasis in these mice[16]. We isolated lung cancer cells from two BLI-negative kidneys that showed no signs of metastasis, yet contained DTCs. We also isolated cancer cells from one case of incipient lung metastasis, and from three individual lung macrometastases that evolved spontaneously (Fig. 4a–b). Cells were uniformly isolated through antibiotic selection for one passage in culture and then subjected to scRNA-seq. This selection step, which was essential because DTCs are extremely rare, likely has some effect on gene expression and is a limitation of our model. Clustering[15] identified 18 mouse metastatic populations (Fig. 4c–d) that show distinct patterns of *SOX2* and *SOX9* expression (Fig. 4e). We then correlated the mean expression of each mouse metastatic cluster with means of the four developmental stages observed in patient-derived tumor cells (type I-P, I-Q, II and III annotated in Fig. 3c; Methods), and visualized these correlations using a bipartite graph (Fig. 4f).

The successive stages of metastatic outbreak in this transplantable mouse model mirrored the developmental progression observed in patient metastatic clusters. H2087-LCC cells isolated as DTCs showed a mixture of $SOX2^{low\text{-}int}$ expressing cells and cells negative for both *SOX2* and *SOX9* (Fig. 4e). DTC clusters all specifically correlated with the transcriptional state of quiescent, adult stem-like cells (type I-Q), except for a single subpopulation that correlated with the proliferating stem-like state (type I-P). Moreover, the incipient colonies and the macrometastases all harbored a minority of quiescent stem-like cells (type I-Q), suggesting that transit through this state may be required to form macrometastases. The incipient metastatic colonies also showed enrichment for $SOX2^{high}$ cells and strong correlation with both the regenerative state (type II) and proliferating stem-like cells (type I-P) whose key transcription factors are partially reactivated during regeneration (Fig. 3c). Conversely, clusters predominantly derived from spontaneous macrometastases showed enrichment of *SOX9*, were correlated with these regenerative and proliferating stem-like states (type II and I-P respectively), and gained transcriptional concordance with the AEP state (type III) (Fig. 4f).

Developmental stage-specific differential immune sensitivity

We observed elevated expression of immune and inflammation response genes in type I stem-like cells, which decays exponentially with progressive lung epithelial development (and concordantly, higher *SOX9* expression) (Fig. 5a) and is significantly depleted in type III ($SOX9^{high}$) AEP metastatic clusters (Extended Data Fig. 8a, U = 352574, p = 2e-39). We have previously shown that metastatic outbreaks are determined, in part, by evasion of NK-cell-mediated surveillance[16] and that antibody-mediated depletion of NK cells increases metastatic outbreak from DTCs in our mouse model of lung cancer metastasis[16]. This suggests that along the developmental progression, metastatic cells might exhibit differential vulnerabilities to the innate immune system. As further evidence of this, we observed that patient primary $SOX9^{high}$ tumors are associated with more NK cell infiltrate, as measured by average NK-cell-specific gene signature expression across 510 TCGA lung adenocarcinoma patients[38]. Conversely, stratification of the same patients by *SOX2* expression reveals

reduced NK cell abundance in *SOX2*^{high} tumors, consistent with stage-specific sensitivities (Extended Data Fig. 9a).

To directly test whether developmental transcriptional programs influence sensitivity to NK cell cytolysis, we co-cultured H2087-LCC metastasis-initiating cells with interleukin 2 (IL2)-activated mouse NK cells (Fig. 5b). Imaging and segmentation of thousands of cells enabled unbiased quantitation of endogenous nuclear SOX2 and SOX9 protein expression as detected by immunofluorescence before and after NK cell co-culture. The presence of NK cells selected for cancer cells expressing SOX9 over cells expressing SOX2 (Fig. 5c). In line with this observation, SOX9 overexpression led to a reduction in NK cell-mediated cytolysis, as measured by annexin V/7-AAD accumulation, whereas SOX2 overexpression had no effect on NK cell-mediated killing (Fig. 5d and Extended Data Fig. 9b). To exclude potential bias related to interspecies target-to-effector cell reactivity, we repeated NK co-culture with cell lines derived from mouse *Kras*^{G12D};*p53*$^{-/-}$ LUAD tumors[39]. In general, SOX9 expression was enriched in highly metastatic derivatives as compared to non-metastatic derivatives (Extended Data Fig. 9c). The metastatic derivative KP482T1 was extremely resistant to NK cytolysis *in vitro* and *in vivo16* and showed dramatic enrichment of endogenous SOX9 nuclear protein levels upon NK cell co-culture (Fig. 5e and Extended Data Fig. 9d). While the non-metastatic derivative KP368T1 was more sensitive to NK cytolysis, SOX9 over-expression alone was not sufficient to render it more resistant to NK cytolysis (data not shown). Collectively, these findings suggest stage-specific differential sensitivity to NK cells, with SOX9^{high} metastasis-initiating stages being more resistant.

To better understand the link between *SOX9* expression and resistance to NK cytolysis, we queried the gene expression of stress ligands recognized by NK cells[40] as well as MHC Class I inhibitory ligands across patient-derived tumor cells assigned to the quiescent stem-like state (type I-Q, associated with latent DTCs), the *SOX2*^{high} regenerative state (type II, associated with incipient metastasis) and the *SOX9*^{high} alveolar epithelial progenitor state (type III, associated with macrometastases). NK activating ligands correlated with proliferation (Spearman $R = 0.70$, $p < 0.001$) and did not segregate by *SOX2* or *SOX*9 expression (Fig. 5f). However, MHC Class I transcripts were dramatically upregulated in both the quiescent stem-like cells (type I-Q) that persist long term *in vivo* and in the *SOX9*^{high} AEPs (type III) that predominate in macrometastases and exhibit NK-resistance *in vitro*. Along this progression, we observed cells that exit the quiescent state while invoking a *SOX*2^{high} regenerative program, downregulating MHC Class I markers of self and upregulating NK activating ligands, rendering them potently susceptible to NK cell clearance (Fig. 5f). Indeed, SOX9 overexpression in metastasis-initiating H2087-LCCs increased the relative expression of canonical HLA genes, especially HLA-B (Fig. 5g), which we validated at the protein level (Fig. 5h). Although the NK activating ligand *RAET1L* increased with *SOX9* expression in patient data and within our H2087-LCC overexpression system (Fig. 5f and Extended Data Fig. 9e), there is strong evidence that NK cell responses are tilted towards inhibition when both activating and inhibitory ligands are engaged[41]. Likewise, the E8.5 mouse embryo shows distinct spatial patterning between *SOX2* and *SOX9* stages (inferred from single-cell expression data; see Methods); here we also observed nearly exclusive expression of MHC Class I markers in *SOX9*-expressing cells (Fig. 5i). Finally, the TCGA cohort of 510 LUAD patients exhibited a significant positive

correlation between MHC Class I and *SOX9* target gene expression (Fig. 5j). Together, our analyses suggest that *SOX9*^{high} cells evade NK cell-mediated clearance by inducing MHC Class I markers of self and that proliferative, *SOX2*^{high} cells are actively constrained by NK cell-mediated surveillance.

To test this hypothesis *in vivo*, we intracardially injected H2087-LCCs into the arterial circulation of athymic mice. Thirty days later, we administrated anti-GM1 antibody to trigger NK cell depletion (Fig. 6a), which facilitates the robust outbreak of macrometastases as evidenced by BLI signal (Extended Data Fig. 10a) and histological analysis[16,42]. To distinguish whether the continuum of developmental states observed in each macrometastasis was derived from a single disseminated cell (evidence of tumor cell plasticity) or from polyclonal seeding, we adapted a lineage-tracing rainbow system that relies on stochastic expression of trichromic reporters (Cerulean, Venus, and mCherry) in single cells[43]. The ratio between the reporters endows each cell and its clonally derived population with a unique spectral property that remains stable after multiple passages. Using both high-resolution fluorescence imaging as well as fluorescence-activated cell sorting (FACS) with this system, we found that metastatic lesions formed after NK cell depletion were largely monoclonal (Extended Data Fig. 10b–c).

Despite their monoclonal origin, analysis of NK cell-depleted macrometastases by scRNA-seq showed that *SOX2*+ and double-negative cell types escape more often than in spontaneously arising metastases (Fig. 6b). Clustering of NK cell-depleted metastases[15] (Extended Data Fig. 10d–e) identified 18 populations that correlated with the four developmental stages observed in patient-derived tumor cells (Methods); most NK cell-depleted clusters showed correlation with the type I (P/Q) and type II states (Extended Data Fig. 10f). Each NK cell-depleted metastasis was rich in type I and II cells, whose growth was otherwise restricted in metastases spontaneously arising in the presence of surveilling NK cells (Fig. 6c). The highest *SOX2*-expressing cluster differentially upregulated the EMT marker vimentin (*VIM)* and *DKK1*, an autocrine WNT inhibitor associated with metastatic latency[16], and downregulated E-cadherin (*CDH1*) (Fig. 6d). Furthermore, this cluster showed reduced expression of multiple MHC Class I molecules, which would render this subpopulation susceptible to killing by NK cells (Fig. 6d). Finally, we adapted a deep learning segmentation method to enumerate the nuclear distribution of SOX2 and SOX9 protein expression in NK cell-depleted versus spontaneous macrometastases evaluated by immunofluorescence in fixed tissues *in situ.* SOX9^{high} cells were enriched only in spontaneous macrometastases that evolve in the presence of NK cells compared to metastases arising after NK depletion (Fig. 6e–f).

DISCUSSION

Understanding tumor cell heterogeneity within a developmental framework is a powerful strategy for identifying functional commonalities across seemingly disparate tumor states, especially in the context of metastasis—a process that must surpass multiple bottlenecks with diverse requirements[10]. Our work demonstrates convergence on common developmental and regenerative processes during cancer progression and presents evidence for developmental stage-specific immune-mediated pruning during metastatic outbreak,

paving the way for future mechanistic dissection and therapeutic targeting of metastatic cancer.

We show that primary LUAD tumors emulate lung epithelial cell stages involved in normal regeneration after lung injury and display stark lineage promiscuity across nearly all cell types of the regenerating alveolar and bronchial lineages. Conversely, human metastases show reduced lineage differentiation and greater stem-like character. Through deeper characterization of tumor subpopulations comprising metastatic lesions, we provide evidence for a developmental continuum spanning the adult stem cell state, intermediate *SOX2*^{high} regenerative subpopulations, as well as proliferative *SOX9*^{high} epithelial lung progenitors. Importantly, we find that these distinct developmental stages exhibit differential sensitivities to NK cell surveillance in patient and mouse models, suggesting that developmental plasticity during metastatic progression is actively sculpted by the immune system. To better understand this stage-specific pruning, it will be important to define the specific target genes of these key *SOX* transcription factors.

By recapitulating these findings in a mouse model of metastatic latency and outbreak, we show that the developmental progression seen in patient samples mirrors the various stages of metastatic outbreak, from the singly disseminated metastasis-initiating cells, to *SOX2*^{high} incipient metastases, and ultimately, to large macrometastases with a preponderance of *SOX9*^{high} cells. Stage-specific pruning in this model is alleviated upon NK cell depletion, which allows *SOX2*^{high} and double-negative cancer cells to emerge alongside *SOX9*^{high} cells as prominent components in metastatic lesions.

Our work shows that tumors can engender phenotypic heterogeneity by subverting tissue repair mechanisms, wherein distinct developmental stage-specific susceptibilities enable context-dependent adaptation and immune evasion. The expansion of DTCs to form metastases is kept in check by NK cells. Targeting these lethal and solitary metastasis-initiating cells will require entirely different strategies than those targeting the AEPs that constitute the bulk of overt macrometastases, whereas macrometastases that evade NK cell-mediated surveillance may be more effectively targeted by therapies harnessing the adaptive immune system.

METHODS

Human Specimens

Non-involved lung, tumor tissues, and metastatic lesions were obtained from patients with lung adenocarcinoma undergoing resection surgery at Memorial Sloan Kettering Cancer Center (MSKCC, New York, NY) after obtaining informed consent. All protocols were reviewed and approved by the Institutional Review Board (IRB) at MSKCC (IRB protocol #14–091). Samples were collected from 14 patients spanning stage I-IV disease. Patient resection site, smoking history, primary lesion size, disease stage, diagnostic pathology, oncogenic mutations identified by targeted exome sequencing using MSK-IMPACT[20], and treatment history are annotated in Extended Data Fig. 1; complete MSK-IMPACT results including oncogenic mutations in the investigational panel are detailed in Supplementary Table 3. Care was taken to collect primary tissue from the tumor core and non-involved lung

distant from the primary tumor as a wedge spanning the distal and conducting airway. Tissue samples were immediately placed in RPMI media (Corning) or Hypothermosol on ice and dissociated using both mechanical and enzymatic digestion (Human Tumor Dissociation Kit #130–095-929, Miltenyi Biotec), generally within 1 h of surgical resection. Tissues were minced with a razor blade in the Miltenyi enzyme mix according to the manufacturer's specifications and transferred to a Gentle MACS Octo Dissociator with heaters (# 30–096-427, 37°C) for further mechanical dissociation. Upon completion, cell suspensions were passed through a 70 μm filter and washed twice with FACS buffer (2% heat-inactivated FBS, 1 mM EDTA and Pen/Strep in PBS without Ca or Mg). Red blood cells were lysed in Red Blood Cell Lysis Solution (ACK lysis buffer) once or twice depending on red blood cell content, and final single-cell suspensions were made in Hanks' Balanced Salt Solution (HBSS). For scRNA-seq, the remaining cell suspensions were subsequently flow sorted with a BD FACSAria II cell sorter fitted with a 100 μM nozzle to enrich for viable, single cells according to forward and side scattering, and DAPI exclusion. Cells were sorted directly into RPMI media with 10% FBS, washed thrice and re-suspend in PBS with 0.04% BSA for single cell encapsulation. Final cell concentrations were determined with a hemocytometer. Three additional matched primary tumor and metastases were acquired under IRB protocol #17–239.

Droplet-based scRNA-seq

Patient Data (10X Genomics Protocol): The 10X Genomics Chromium Platform was used to generate a targeted 5000 single cell Gel Bead-In-Emulsions (GEMs) per sample, loaded with an initial cell viability of ~90%. scRNA-seq libraries were prepared following the 10X Genomics user guide (Single Cell 3′ V2 Reagent Kits User Guide PN-120233, 10X Genomics). After encapsulation, emulsions were transferred to a thermal cycler for reverse transcription (RT) at 53°C for 45 min, followed by heat inactivation for 5 min at 85°C. cDNA from the RT reaction was purified using DynaBeads MyOne Silane Beads (Thermo Fisher Scientific) and amplified for 12 cycles using Amplification mix and primers provided in the Single Cell 3' reagents module 1 (10X Genomics). After purification with 0.6X SPRIselect beads (Beckman Coulter), cDNA quality and yield was evaluated using Agilent Bioanalyzer 2100. Using a fragmentation enzyme blend (10X Genomics) the libraries were fragmented, end-repaired and A-tailed. Products were double side cleaned using 0.6X and 0.8X SPRIselect beads, and adaptors provided in the kit were ligated for 15 min at 30°C. After cleaning ligation products, libraries were amplified and indexed with unique sample index i7 through PCR amplification. The number of PCR cycles was chosen based on cDNA yield for each sample individually. Final libraries were double-side cleaned using 0.6X and 0.8X SPRIselect beads and their quality and size was evaluated using Agilent Bioanalyzer 2100. Libraries were sequenced by pooling two per lane on HiSeq2500 (Illumina) paired-end read flow cell with a 10.5pM loading concentration, sequenced 26 cycles on the forward read (10X Barcode + UMI), followed by 8bp I7 Index (Sample Index) and 98bp on the reverse read.

Mouse Data (inDrops Protocol[11,12]): Barcoded hydrogel bead synthesis, single cell encapsulation and library construction were done following a modified inDrops protocol[12,13]. Barcoded hydrogel beads were synthesized in house and contained a poly(T)

sequence, T7 promoter, PE1 sequencing primer, unique cell barcode, unique molecular identifier (UMI) and a photo-cleavable linker. Cells were re-suspended in PBS with 0.04% BSA at ~400 cells/μL. Before encapsulation, cells were mixed at a 1:1 ratio with OptiPrep mix (32% OptiPrep Density Gradient Medium (Sigma-Aldrich), 0.08% BSA in PBS). Single cells were encapsulated in droplets together with barcoded hydrogel beads and reverse transcription mix (SuperScript™ III Reverse Transcriptase, SUPERase-In, FS buffer (Thermo Fisher Scientific) NP-40, DTT, $MgCl_2$, Tris and dNTPs) using a microfluidic device with three inlet channels for aqueous solutions (barcoded hydrogel beads, RT mix and cells), an oil inlet, and an outlet channel for emulsion. Emulsion was collected for 30 min in a 1.5 mL tube in ice block. Immediately after encapsulation, primers were released from hydrogel beads by incubating the emulsion on ice under UV light for 7 min. The emulsion was then transferred to heat block at 60°C for 1 min, 50°C for 2 h (RT), followed by heat inactivation at 70°C for 15 min. The emulsion was then divided into smaller aliquots that contained an estimated 5000 cells. A few drops of 20% 1H,1H,2H,2H-perfluorooctanol (PFO, Sigma-Aldritch) was added on top of each tube to break emulsions, which were then transferred and stored at −80°C. Hydrogel beads were removed by filtration through a spin column (Zymo Research) and excess primers were digested using Exo1, HinF1 and FastAP enzyme mix (Themo Fisher Scientific). After the DNA/RNA duplex was purified using 1.2X SPRIselect beads (Beckman Coulter), second strand synthesis (SSS) was done using NEBNext® Ultra™ II Non-Directional RNA Second Strand Synthesis Module (New England BioLabs) at 16°C for 2.5 h followed by inactivation at 65°C for 20 min. SSS reaction material was then amplified using HiScribe™ T7 High Yield RNA Synthesis Kit (New England BioLabs) for 15 h at 37°C. Reaction products were purified using 1.2X SPRIselect beads (Beckman Coulter) and their quality was evaluated on Agilent Bioanalyzer 2100. Amplified material was fragmented using RNA fragmentation reagents (Ambion Life Technologies) at 70°C for 2.5 min. Fragmentation was stopped by addition of ice cold 1.2X SPRIselect beads mixed with STOP solution. After purification, fragmented aRNA was mixed with PE2-STUB (IDT) which contains random hexamers, incubated for 3 min at 70°C, cooled on ice, and reverse transcribed using PrimeScript RTase (Takara Bio USA) for 60 min at 42°C. Libraries were purified with 1.2X SPRIselect beads (Beckman Coulter) and amplified via PCR (Kapa 2× HiFi HotStart PCR mix, Kapa Biosystems) using P1-P2 Illumina index primers; optimal cycle number was determined using qPCR. Amplified and indexed libraries were cleaned two times using SPRIselect double-sided size selection (0.6X and 0.8X). Library size was analyzed using Agilent Bioanalyzer 2100 and quantified by Qubit dsDNA HS Assay kit (Thermo Fisher Scientific). Libraries were sequenced one per lane of HiSeq4000 (Illumina) paired-end read flow cell, loaded at a 10.5pM concentration with 15% PhiX spike-in. 54 bp were sequenced in the forward read (inDrop Barcode + UMI) and 46 bp on the reverse read.

scRNA-seq Computational Analysis

Pre-processing, cell selection and filtering of patient droplet-based scRNA-seq data: The Sequence Quality Control (SEQC) package[13] was utilized to process the data, constructing a count matrix from raw reads, including de-multiplexing, alignment, error-correction, and the generation of a raw digital expression matrix by collapsing groups of reads with the same unique molecular identifier (UMI), cell barcode and gene annotation.

Alignment to the hg38 annotation was restricted to transcribed, polyadenylated RNA of length > 200 nucleotides (gene biotypes accessible by 3' mRNA sequencing technologies) to increase mapping specificity. SEQC then follows with a number of filtering steps to ensure data quality. Viable cells were distinguished from droplets consisting of ambient mRNA transcripts arising in solution due to premature lysis or cell death based on library size; whereby cells were filtered beyond the knee point of the second derivative of the empirical cumulative density function of total cell transcript counts (Extended Data Fig. 2a). Additionally, cells with low complexity libraries (in which detected transcripts are aligned to a small subset of genes) were filtered (Extended Data Fig. 2b). Cells with > 20% of transcripts derived from mitochondria were considered apoptotic and also excluded (Extended Data Fig. 2c). After clustering (described below), two minority cell clusters were additionally excluded from patient data with characteristically low library size and signatures of partial cell lysis that escaped the automated filters described above. Genes detected in fewer than 10 cells or genes with low expression levels, identified as those with count values < 5 standard deviations from the second mode of the log-log distribution of total transcript counts/gene, were also excluded. This yielded a total of 40,505 patient-derived cells with a median library size of 4,038 transcripts per cell (Extended Data Fig. 2d–e), for downstream analysis (Fig. 1–3 & 5 Data).

Normalization and imputation: The filtered count matrix was normalized for library size per cell, whereby the expression level of each gene was divided by the cell's total library size and then scaled by the median library size of all cells. Principal Component Analysis (PCA) was computed using randomized principal component analysis[48] applied to the normalized count matrix. Finally, MAGIC imputation[14] was applied to the median-normalized count matrix to further denoise and recover missing gene values using conservative parameters (t = 3–4, k = 27–28). Imputation was performed using the first 20–40 principal components of the normalized count matrix. The number of principal components was selected per dataset based on the knee point of the cumulative explained variance[49] and accounted for >90% of variance in patient data (acquired on 10X platform) and ~60% of variance in mouse data (acquired on inDrop platform). Within the *epithelium/stroma* and *lymphoid* subsets, ordinary least squares was applied to linear regress library size out of each principal component prior to MAGIC imputation because a partial correlation was observed between some principal components and library size. When incorporating metastatic samples in patient data and to facilitate their comparison to the primary samples, the MAGIC imputed data was secondarily normalized by dividing the imputed expression level of each gene in a cell by the cell's total imputed library size and then scaling by the median imputed library size of all cells. Normalization post-imputation was necessary for this dataset because metastases were transcriptionally more active, with larger and more complex libraries when compared to epithelium of the normal lung or primary tumors. In all cases, imputed data was used for data visualization, clustering and to explore trends of individual genes or gene-gene correlations in the data; it was never used to identify differentially expressed genes or enriched pathways.

Data visualization: The global atlas of all patient cells, including diverse *epithelium/stroma, myeloid* and *lymphoid* cell subpopulations (Fig. 1c, Extended Data Fig. 2f–g), and

all immune subsets (Extended Data Fig. 3–4), were visualized using the Barnes-hut approximate version of t-SNE[50] (https://github.com/lvdmaaten/bhtsne) computed on the principal components of the imputed data. This visualization was appropriate given the diversity of cell types represented in these data subsets. Force-directed graphs[51] were alternatively used to visualize epithelial and tumor subsets, which better represent cell state transitions and local relationships between cancer cells and normal epithelial subpopulations, while maintaining a coherent global structure (Fig. 2 and 4, Extended Data Fig. 6–8 & 10). Force-directed layouts were computed on the principal components of the imputed data using a k-NN graph[15]; whereby the adjacency matrix representing this k-NN graph was converted to an affinity matrix using an anisotropic Gaussian kernel[14] because the standard Gaussian kernel does not account for large differences in densities in the data. The adaptive kernel used distance to the k = 10 neighbor as the scaling factor for each cell to determine the affinities[14]. The affinity matrix was then used as input to compute the force-directed layout using the ForceAtlas2 python module[51]. For both visualization methods, the number of principal components was selected per dataset based on the knee point of cumulative explained variance. Post-imputation, no more than 20 principal components explained > 90% of variance in all patient data (acquired on 10X platform) and no more than 32 principal components explained > 80% of variance in mouse data (acquired on inDrop platform).

Customized plotting functions: All visualizations were performed in Python and are demonstrated in a Jupyter notebook available for download (see Code Availability). While most figures were generated using routine plotting functions, herein we elaborate on three customized functions utilized throughout the paper: dot plots, 2D kernel density estimate (KDE) distributions, and the construction of bipartite graphs.

Dot plots: For each gene and categorical grouping (i.e. cell type) a dot is plotted. Each dot represents two values: the fraction of cells expressing a given gene in each category (visualized by the size of the dot) and the average expression of expressing cells within each category (visualized by color). A gene is considered expressed if its normalized (un-imputed) expression is greater than zero. Within each category the normalized (un-imputed) gene expression is averaged only over cells expressing the given gene.

KDE Visualizations: The cell density distribution across two variables (i.e. genes, fraction of lineage-specific markers) was visualized in two dimensions using the bivariate kernel density estimate function in Seaborn. We used default parameters (Gaussian kernel) and specified 10 contour levels. The lowest contour of the bivariate KDE plot was not shaded when plotting multiple cell densities on the same axis.

Bipartite graphs: To visualize linkage between two disjoint and independent categorical sets we used NetworkX to construct a weighted bipartite graph. Here, nodes correspond to each categorical variable (i.e. mouse metastatic clusters and developmental states observed in patient tumor cells) and edges connect nodes across independent sets. Edges are weighted by variables linking the two independent sets like their genome-wide correlation (i.e. Fig. 4f and Extended Data Fig. 10f) or number of co-occurring cells (Supplementary Data Fig. 1b).

Edges are subsequently filtered by criteria specified in each figure legend (i.e. Pearons R > 0.20 and p < 0.05). Nodes of the resulting weighted graph are positioned according to the Fruchterman-Reingold force-directed algorithm implemented in NetworkX using default parameters.

Meta-cell type annotation: Given the large variation observed in cell size as represented by total mRNA abundance in unsorted patient tumors, all cells were initially assigned to one of three meta-cell classes: *lymphoid*, *myeloid* and *epithelial/stromal* cell types, on a per-cell basis according to their expression of canonical immune, myeloid, mesenchymal and epithelial cell markers manually curated and listed in Supplementary Table 2. The cumulative expression of imputed counts per canonical marker list was computed and then z-normalized across cells. Cells were hierarchically clustered by these z-normalized scores using the cosine distance metric (Extended Data Fig. 2h). Meta-cell type assignments were subsequently smoothed by a nearest-neighbor vote. The purpose of this annotation was to simply isolate lymphoid cells, which showed reduced transcript capture rates, as expected[13] (Extended Data Fig. 2g). Cell type annotations were subsequently performed separately within the *lymphoid* compartment and the *myeloid/epithelial/stroma* compartment using refined, graph-based methods to avoid introducing biases from cell type-specific capture rates.

Phenograph clustering within meta-cell types: *Myeloid/epithelial/stromal* cells (Extended Data Fig. 3) and *lymphoid* cells (Extended Data Fig. 4) were directly subset from the imputed count matrix described above. Phenograph clustering[15] was computed directly on the imputed count matrix of each subset and 39 phenotypic cell types were identified within the *myeloid/epithelial/stromal* compartment (parameter k = 30, Extended Data Fig. 3a) and 21 phenotypic cell types were identified within the *lymphoid* compartment (parameter k = 50, Extended Data Fig. 4a). For the latter, the parameter *k* was chosen to be consistent with the larger size of the lymphoid subset.

Cluster based differential gene expression analysis: Cell type-specific gene signatures were identified by ranking differentially expressed genes (DEG) between each Phenograph cluster and all other clusters using the R package MAST[44]. MAST was run using default parameters with normalized counts (without imputation) as the input and the Bonferroni correction was applied to correct for multiple hypothesis testing (p_{adj}). DEGs were reported per cluster according to their fold change and p_{adj} value. DEG were ranked for GSEA according to: $rank = -10 * log_{10}(p_{adj}) * sgn[log_2(fold\ change)]$. A subset of 784 housekeeping genes related to translation and ribosomal RNA transcription and processing (listed in Supplementary Table 2) were excluded from ranked DEGs prior to Gene Set Enrichment Analysis (GSEA).

Gene signatures for cluster annotation: A custom annotation file was generated to probe the role of normal lung epithelial development and regeneration in cancer. This was assembled by integrating gene signatures of lung epithelial cell types identified by single cell sequencing in the developing mouse (D18.5) [5,23–25] with gene ontology (GO) and Reactome molecular signatures containing key words related to normal lung development:

Wnt, TGFβ, sonic hedgehog, Notch, retinoic acid, Hippo, FGF, development or wound response[52], as well as the complete Hallmark Gene sets (see Supplementary Dataset). Seven defined epithelial cell types were ultimately annotated in this paper and their associated lineage-specific gene-sets are listed in Supplementary Table 1. These were sourced from LungGENS[23,24] (https://research.cchmc.org/pbge/lunggens/) for all mature lung epithelial lineages (AEC1, AEC2, club and ciliated) because these gene sets encompassed and expanded upon signatures defined in earlier studies[25]. Basal cells were defined by genes reported as differentially expressed ($p < 0.001$ and fold change > 2.8) in KRT5-GFP^{high} and Lectin^{+} tracheal epithelial cells sorted from transgenic mice[53]. AEPs were likewise defined by reported genes differentially expressed ($p < 0.01$, fold change > 2) in human cells sorted from the normal lung using human-AT2-specific HTII-280 antibody and surface markers TM4SF1 and EPCAM, as compared to other human AEC2 cells[5]. A short list of manually curated, canonical markers expressed by neuroendocrine *cells (CALCA, ASCL1, PPIG, DYNLRB1, CGA, ASH1L and DDC) and cells producing mucins (AGR2, TFF1, TFF3, PARM1, ADGRE2, GCNT3* and *MUC* variants) were additionally included as independent cell types in the lineage-specific gene sets (Supplementary Table 1). Published GMT files related to stromal cell types defined by single cell sequencing in the developing mouse lung[5,23–25] and molecular signatures of immune cell types[54,55] were additionally queried by GSEA when evaluating the merged myeloid, epithelial and stromal cell types in Extended Data Fig. 3.

To visualize GSEA results in an intuitive manner for cell type annotation, gene signatures distinguished by a Bonferroni corrected $p_{adj} < 0.05$ and an absolute Normalized Enrichment Score $abs(NES) > 1.5$ were considered significant and hierarchically clustered (clusters and gene signatures) by NES according to the Euclidean or cosine distance metric. The cosine distance metric was chosen when clustering highly diverse cell types (myeloid, epithelial and stroma in Extended Data Fig. 3b), so that instances of gene signatures drive clustering independent of their magnitude. When probing cell state heterogeneity within the epithelial compartment, the Euclidean distance metric was utilized, although clustering in this instance was largely independent of the distance metric. Values not meeting these criteria were whited out on the heatmap. For all epithelial data subsets interrogated in this paper, we additionally report an unfiltered list of all gene signatures distinguished by $p_{adj} < 0.25$. Columns of NES heatmaps are labeled by Phenograph cluster and by cell type when annotated.

Selection of variably expressed genes in single cell data: Variably expressed genes were selected per single cell dataset based on the dispersion of their expression across all cells. The distribution of the $\log_{10}$ mean and variance per gene across all cells, for all genes, was fit to a polynomial using least squares minimization. Highly variable genes were identified as those whose variance was greater than the mean fit.

Cell type annotation: The NES for gene signatures characterized by $abs(NES) > 2$ and $p_{adj} < 0.05$ within the myeloid, epithelial and stromal meta-subset, clustered according to the cosine distance metric (Extended Data Fig. 3b) show clear segregation between epithelial, stromal and myeloid cell types. DEGs per Phenograph cluster ranked according to p_{adj} and

fold change are ranked in Supplementary Table 4. However, GSEA results can be promiscuous and difficult to interpret. Therefore, cell type assignments were further informed by examining the correspondence between Phenograph cluster medians and the expression profiles of sorted immune populations, for which bulk microarray and RNA-seq datasets were available[56,57]. First, bulk microarray datasets were log2 transformed and library-size normalized. Then, the mean expression per cell type was computed across biological replicates and centered by mean subtraction per gene across all cell types. Correspondingly, the median imputed expression of each gene per Phenograph cluster was computed within the *myeloid/epithelial/stroma* and *lymphoid* subsets, and likewise centered by subtracting the mean expression level per gene across all clusters. Finally, the Pearson correlation between the centered transcriptional profile of each Phenograph cluster and bulk immune cell type was computed in a pairwise manner using all genes variably expressed in single cell data (described above) and detected in bulk immune data (~6000 genes). For each pairwise correlation, the Python package scipy was utilized to compute a *p* value testing for non-correlation. Hierarchical clustering of the Pearson correlation between each Phenograph cluster and bulk immune cell type is shown in Extended Data Fig. 3c and Extended Data Fig. 4b for the *myeloid/epithelial/stroma* or *lymphoid* subsets respectively; where each row is colored by Phenograph cluster. Only correlation coefficients characterized by $p < 0.01$ are visualized; all others are whited out. As expected, Phenograph clusters annotated as epithelium or stroma in Extended Data Fig. 3c do not show strong correlation with immune signatures. In agreement with GSEA, myeloid cell types positively correlated with monocyte, macrophage and dendritic cell signatures as shown in Extended Data Fig. 3c. Distinct subset of mast cells were also identified that did not associate significantly with any pathways queried by GSEA.

Bulk mRNA signatures do not exist for many epithelial and stromal cell types in the lung; however, a growing database of cell type-specific gene signatures in the developing mouse lung have now been annotated by LungGENS (https://research.cchmc.org/pbge/lunggens/)[23,24]. Mouse genes were converted to human genes by capitalization. Several of these signatures showed enrichment with Phenograph clusters within the *myeloid/epithelial/stroma* subset by GSEA (Extended Data Fig. 3b).

All together, final cell type assignments were assembled based on GSEA results and correlation with published immune transcriptional profiles. Final epithelial, stromal and myeloid cell type assignments are shown in Extended Data Fig. 3d, which were further confirmed by evaluating the imputed expression of their canonical cell type markers (Extended Data Fig. 3e). For example, epithelial cells abundantly expressed E-cadherin (*CDH1*) as well as the lineage-determining transcription factors *SOX2* and *SOX9*. The intermediate filament protein, vimentin, predominantly marked fibroblast-like cells positive for α-smooth muscle actin and desmin, another intermediate filament associated with contractile cells throughout the body. A minority of epithelial-like cells and a majority of macrophages also expressed increased levels of vimentin, as previously reported in alveolar macrophages[58]. Some mesenchymal cell types expressed platelet-derived growth factor receptor α (*PDGFRα*) and *LGR5*, characteristically expressed in the Wnt-responsive alveolar niche[59,60]; whereas others expressed *PDGFRβ* and neural/glial antigen 2, encoded by chondroitin sulfate proteoglycan *4*, *CSPG4*, characteristic of pericytes. *S100A4* (also

known as fibroblast-specific protein 1), a critical mediator of fibrotic progression[61], was additionally used as a marker of fibroblasts in the lung[62–64]. Likewise, *S100A4* was expressed in some immune cells, where it has been shown to play a role in macrophage recruitment and chemotaxis in vivo[65]. Pulmonary endothelial cells were distinguished from other stromal cell types by their expression of vascular endothelial (VE)-cadherin (*CDH5*) and the early endothelial cell marker *CD34*. Within the myeloid compartment, antigen-presenting macrophages and dendritic cell type assignments were validated by their expression of MHC class II antigens, as well as other canonical markers including *CD40*, the chemokine receptor *CCR7*, and *MARCO* (*macrophage receptor with collagenous structure*). A subset of *ITGAM(CD11B)+* immature myeloid cells expressing both *CSF1R* and immunosuppressive *IL10* were further distinguished as myeloid-derived suppressor cells (MDSCs), which have been shown to dampen inflammatory responses[66,67]. These cell type assignments (Extended Data Fig. 3d) were mapped back to the complete patient dataset in (Fig. 1c) using the same color scheme.

In the *lymphoid* subset, Phenograph clusters largely correlated with distinct B, cytotoxic NK, NKT and T cells (Extended Data Fig. 4b). Cell type assignments within the B, NK and T cells of the *lymphoid* compartment were likewise refined by examining the median imputed expression of canonical marker genes on a per cluster basis (Extended Data Fig. 4c). Imputed expression levels of individual genes were z-normalized by subtracting the mean and dividing by the standard deviation per gene across all cells. DEGs per Phenograph cluster, ranked according to their Bonferroni-corrected *p* value and fold change, are annotated in Supplementary Table 5. Final lymphoid cell type assignments are shown in Extended Data Fig. 4d, which were further confirmed by evaluation of canonical cell type markers (Extended Data Fig. 4e–f). Regulatory B cells and four subpopulations of mature plasma cells distinguished by specificity of their immunoglobulin secretion were observed. Distinct T and Natural Killer (NK) lymphoid cell types were identified that corresponded to known immune subpopulations, like *CD4+IL2RA(CD25)+FOXP3+* regulatory T cells. NK cells and to a lesser extent, NKT cells, most abundantly expressed the surface receptors *NKG7, NCR3, FCGR3A(CD16)* and *NCAM1(CD56)*. NKT cells were distinguished from NK cells by specific expression of *CD3*. Again, these cell type assignments (Extended Data Fig. 4d) were mapped back to the complete patient dataset in (Fig. 1c) using the same color scheme.

Final, merged cell type assignments are shown in Fig. 1c, which are further validated based on the imputed expression of their canonical cell type markers in Extended Data Fig. 5a. We observe most cell types are detected across all patient samples, but display different amounts of mixing between patient samples (Extended Data Fig. 2f and Extended Data Fig. 5a).

Mixing of samples in cell types: We wished to evaluate which cell types were well represented across all patients and which cell types were more patient specific in their nature. To quantify the degree of mixing between patients within each cell type, we used an entropy-based metric[13], along with bootstrapping to correct for cell type size (which ranged from 89 to 7254 cells), such that we uniformly sampled 100 cells from each cell type and computed the distribution of patients across these cells. We then computed the Shannon entropy for this distribution across patient samples m = 1,…,17 according to

$H_i = -\sum_{m=1}^{17} p_i^m \ln(p_i^m)$, where P_i represents the fraction of cells in patient i. We repeat this procedure 100 times for each cell type, and visualize the entropy distribution per cell type using a kernel density plot (Extended Data Fig. 5b). We repeated this analysis for cell states observed in analysis of epithelial lineages derived from the normal lung, primary tumor and metastases (Fig. 2c and Extended Data Fig. 8e). High entropy indicates that a cell state is highly reproducible across patient samples; whereas low entropy indicates that the cluster/cell state is mostly derived from the same sample and are patient-specific. We observe no patient-specific clusters across the cell types annotated in our global cell atlas; in fact, all cell types are detected in at least four patients with the majority detected in nearly all 17 patient samples (Extended Data Fig. 5a–b). As also observed in [13], intra-tumor macrophages show the highest degree of patient specificity.

Individual sample normalization, imputation and clustering: Merging data across patients increases statistical power for detecting cell types and is widely used in single cell analyses[13,68,69]. To test the robustness of our ability to detect cell types when analyzing patients individually, we analyzed the cellular composition of each patient one-by-one and report the patient frequency per cell type (Extended Data Fig. 5d–e). We processed each patient from raw cell counts using the cohort level analysis pipeline described above, including only cells from each individual patient for library size normalization, PCA, MAGIC imputation, Phenograph clustering and tSNE projection. Contrary to the cohort analysis, we did not separate the lymphoid and non-lymphoid cells before clustering all populations within each patient. Moreover, due to the smaller number of cells, Phenograph was applied to a k-NN graph of only 20 nearest neighbors. Cell type annotations of these clusters followed the same procedures used for the cohort level analysis. For all major cell types, patient frequency was concordant across individual and grouped annotations (Extended Data Fig. 5a,e).

Lineage annotation within epithelial cells: Our analysis of the epithelial compartment focused on the relationship between primary tumor cells and mature or regenerative epithelial lineages of the post-natal lung; therefore we evaluated cell states within the epithelial compartment in a step-wise fashion: (E1) *the normal, adult human lung* (Extended Data Fig. 6), (E2) *the merged normal lung and primary tumor* (Fig. 2, Extended Data Fig. 7) and (E3) *the merged normal lung, primary tumor and metastases* (Fig. 3, Extended Data Fig. 8). Normalization, imputation, visualization and DEG per Phenograph cluster were computed for each epithelial subset (E1–3) separately, as described above, from the level of raw counts. Parameters that varied across processing of the three epithelial subsets, including the number of principal components (PCs) utilized for MAGIC imputation, visualization and clustering, and the number of nearest neighbors, *k*, used to construct the k-NN graph for Phenograph clustering, are annotated below.

(E1) *Normal, adult human lung*

20 PCs, > 95% explained variance, Phenograph $k = 140$, number of Phenograph Clusters = 4

(E2) *Normal lung and primary tumour epithelium*

20 PCs, > 95% explained variance, Phenograph $k = 50$, number of Phenograph Clusters = 15

(E3) *Normal lung, primary tumour and metastatic epithelium*

14 PCs, > 95% explained variance, Phenograph $k = 25$, number of Phenograph Clusters = 25

In all cases, the knee point method was utilized to select the number of principal components for each processing step (< 20 components explained > 95% of variance in all data subsets). Phenograph clustering was computed on the principal components of the imputed count matrix. The number of nearest neighbors, *k*, used to construct the k-NN graph was selected such that the Jaccard graph per data subset was fully connected ($k = 6–8$) and it was within a range of *k* for which cluster assignments were robust. Robustness of Phenograph clusters was evaluated by computing the adjusted rand index (ARI)[15] between categorical cluster assignments made using all pairwise values of *k*. For each pairwise comparison, we compute the ARI across all cells. The resulting *k*-by-*k* matrix of ARI values was visualized as a two-dimensional heatmap and *k* was selected within a range that showed stable concordance between categorical assignments (ARI > 0.75). For each epithelial subset, Phenograph clustering was performed using a value of *k* for which the graph was fully connected and the ARI > 0.75.

(E1) Epithelial cell types detected in the normal, adult human lung: A total of 658 cells assigned to the epithelial clusters defined in Fig. 1c and sampled from normal lung were subset from the count matrix to define normal epithelial lineages in the adult, human lung. Normalization, imputation, visualization and DEG per Phenograph cluster were computed for this data subset separately as described above. Phenograph cluster assignments were stable for higher values of *k*, as determined by the ARI above, and four clusters were identified in the normal lung sampled from four patients, which are annotated by cell type in Extended Data Fig. 6a–b. Cell type annotations were informed by GSEA using the Supplementary Dataset and by evaluating the imputed expression of lineage-specific genes listed in Supplementary Table 1. A complete list of DEG per cell type ranked according to their Bonferroni-corrected *p* value and fold change are listed in Supplementary Table 6. Hierarchical clustering of the imputed expression of the top 60 DEGs per cell type (distance metric = cosine), z-score normalized per gene across cells, clearly segregate the four epithelial lineages. Canonical lineage-specific genes[25] (Supplementary Table 1) are labeled on the x-axis of this clustered heatmap (Extended Data Fig. 6c) and rows of the clustered heatmap are colored by lineage assignment. For visualization, GSEA pathways distinguished by $abs(NES) > 2.5$ and $p_{adj} < 0.05$ were clustered according to the Euclidean distance metric (Extended Data Fig. 6d) and show specific distal (AEC1 and AEC2) and proximal (club and ciliated) epithelial lineages; white areas of the heatmap indicate pathways that did not meet this criteria. An unfiltered list of all gene signatures distinguished by $p_{adj} < 0.25$ is provided in Supplementary Table 7. Cell type annotations were orthogonally validated by evaluating the fraction of cells per Phenograph cluster that abundantly express (at or above the 75th percentile of the population expression level) more than 60% of genes within the lineage-specific signature[25] (Extended Data Fig. 6e, Supplementary Table 1). Club cells fail this specific test, despite showing significant enrichment of two Club cell signatures independently derived in the developing mouse embryo at D18.5 (LungGens[23,24] NES =

2.64, p_{adj}=0.01 and Treutlein et al[25], NES = 2.56, p_{adj}=0.02) (Extended Data Fig. 6d). This suggests their annotation may be driven by high expression of few canonical markers including *SCGB1A1, SCGB3A2* and *CYP2F1*, as revealed by downstream analysis in Extended Data Fig. 7a.

(E2) The relationship between primary tumor and normal lung epithelial cell types: We subsequently analyzed the merged 2,140 epithelial cells sampled from both the non-tumor-involved lung and primary LUADs. Again, we subset epithelial cells at the level of the raw count matrix and repeated normalization, imputation, visualization and DEG per Phenograph cluster for this data subset separately as described above. In addition to the four epithelial cell types previously identified in the non-tumor involved lung, eleven more phenotypic states were identified by Phenograph clustering. DEGs per Phenograph cluster ranked according to their Bonferroni-corrected *p* value and fold change are annotated in Supplementary Table 8. DEGs that intersect with lineage-specific genes listed in Supplementary Table 1 and characterized by an absolute fold change > 1.5 and $p_{adj} < 0.05$ are colored by their associated lineage on volcano plots for annotated Phenograph clusters (Fig. 2c and Extended Data Fig. 7a). Ranked gene lists per Phenograph cluster were queried by GSEA using the custom lung development annotation file (Supplementary Dataset). The NES of gene sets characterized by *abs(NES)* > 1.5 and $p_{adj} < 0.05$ are visualized on a heatmap, clustered by Phenograph clusters (columns) and gene signatures (rows) according to the Euclidean distance metric (Extended Data Fig. 7b). White areas of the heatmap show signatures that did not meet the NES and p_{adj} criteria described above. Columns are colored above by cell type annotation; mapping between Phenograph clusters and cell type annotations was achieved by evaluating the lineage-labeled volcano plots and GSEA results. An unfiltered list of all gene signatures distinguished by $p_{adj} < 0.25$ is provided in Supplementary Table 9. Cell type annotations were further supported by waterfall plots of the imputed expression of lineage-specific genes (Extended Data Fig. 7c) within each cell type; cells annotated as mixed lineage (grey) indeed express canonical markers of multiple proximal and distal cell types.

(E3) Cell state heterogeneity within normal, primary tumor and metastases: Phenotypic heterogeneity was evaluated within all 3,786 epithelial cells sampled from normal lung, primary tumor and metastases, subset directly from the raw count matrix (Extended Data Fig. 8a–c). Normalization, imputation, secondary normalization post-imputation, visualization and DEG per Phenograph cluster were computed as described above for this data subset separately. Tissue source (Extended Data Fig. 8b, *Left*) and Phenograph cluster (Extended Data Fig. 8b, *Center*) are visualized on a force-directed layout of all epithelium. For orientation, cell type annotations made during analysis of the normal and primary tumor epithelium (Fig. 2b) were mapped onto this merged dataset; matching between cells across datasets was achieved using their cell barcodes and are labeled by text on the force directed layout (Extended Data Fig. 8b, *Right*). DEGs per Phenograph cluster, ranked according to their Bonferroni-corrected *p* value and fold change, are annotated in Supplementary Table 10. Differentially expressed pathways per Phenograph cluster distinguished by *abs(NES)* > 1.5 and $p_{adj} < 0.05$ were clustered according to the Euclidean distance metric for visualization (Extended Data Fig. 8a). An unfiltered list of all gene signatures distinguished

by $p_{adj} < 0.25$ is provided in Supplementary Table 11. The clustered heatmap is colored along columns by Phenograph cluster (lower) and fraction of tissue source detected per cluster (upper). In the same order, box plots show the fraction of each Phenograph cluster detected per metastasis (n = 5, Extended Data Fig. 8d). Finally, we evaluated the entropy of the distribution of patients in each Phenograph cluster, with bootstrapping to correct for number of cells in each cluster as described above (Extended Data Fig. 8e). Patient-specific cell states are observed in neo-adjuvantly treated primary tumors and metastases (Extended Data Fig. 8e). Interestingly, these are almost exclusively associated with the $SOX9^{+}$ alveolar epithelial progenitor state (type III). Clusters associated with the adult stem to regenerative state (type I-II) are reproducibly observed across multiple patients.

Robustness of clustering to imputation: Imputed data was never used to identify differentially expressed genes or pathways in the analyses, however clustering was performed on imputed data. Therefore, to evaluate the robustness of cluster structure to imputation within the critical analysis of normal lung and primary tumor derived epithelium, we independently clustered the data using PCA applied to the normalized un-imputed count matrix or the normalized imputed count matrix, yielding two sets of categorical assignments hereafter referred to as Imputed Phenograph Clusters (IPC) and Un-Imputed Phenograph Clusters (UIPC). In both cases, the knee point method was utilized to select the number of principal components input into Phenograph using default parameters (k-NN = 50). To compare the probabilities of individual cells co-clustering across Imputed and Un-Imputed Phenograph Clusters, we calculated the normalized mutual information (NMI = 0.81) between these two sets of categorical assignments according to $NMI(IPC, UIPC) = \frac{2\ x\ MI(IPC, UIPC)}{[H(IPC) + H(UIPC)]}$, where MI is mutual information and H is Shannon entropy. Quality of clustering was evaluated between un-imputed and imputed data with normalized mutual information because this facilitates comparison between sets that have different numbers of clusters (imputation increases cluster resolution). Next, we computed the cell-cell co-occurrence matrix for all cells across the Un-imputed vs. Imputed Phenograph clusters (Supplementary Fig. 1a). The row of each imputed Phenograph cluster is colored by its annotated cell lineage (Fig. 2b). We construct a bipartite graph where each cluster is represented by a node, whose diameter is proportional to cell number, and we add an edge between un-imputed and imputed Phenograph clusters sharing more than 50 cells (Supplementary Fig. 1b). The majority of assignments map one-to-one across imputed and un-imputed Phenograph clusters; with imputation sometimes increasing cluster resolution within a given lineage. Importantly, there is negligible mixing between clusters assigned to different lineages with or without imputation, as best visualized by the bipartite graph representation of this cell co-occurrence matrix.

Lineage promiscuity in single cells: Analysis of the combined normal and primary tumor epithelium revealed six Phenograph clusters that differentially expressed canonical markers associated with multiple proximal and distal lung epithelial cell lineages (Fig. 2b–e). DEGs per cluster and lineage-specific gene expression was exclusively analyzed in un-imputed data to ensure that imputation did not artificially introduce unexpected mixing of marker genes. To test whether this lineage promiscuity was also observed at the level of individual cells, we generated a binary matrix representing the single cell expression of specific lineage

markers shown in Fig. 2c. A marker was considered abundantly expressed in a cell if its normalized (un-imputed) counts were in the top quartile of expression across all cells evaluated per gene. The binary heatmap revealed clusters of tumor cells aberrantly expressing markers associated with AEC1, AEC2, club and neuroendocrine lineages at the level of individual cells (data not shown, see Jupyter notebook). Lineage marker combinations also reflected underlying Phenograph cluster structure; for example Phenograph Clusters 0 and 1 showed striking mixing between AEC1 and AEC2 lineages. Next, we compute what fraction of abundantly expressed lineage-specific genes belong to each annotated epithelial cell type. Finally, we visualize the density of cells expressing pairwise fractions of cell-type specific markers, as shown for AEC1 and AEC2 markers (Fig. 2f) for normal epithelial cells assigned to these two lineages, and for mixed-lineage Phenograph Clusters 0 and 1, which show particular promiscuity between these two lineages.

Defining lineage phenotypic volume: Intrigued by the lineage promiscuity observed in primary tumors, we wanted to quantify the degree of tumor cell-intrinsic heterogeneity, relative to epithelial of the normal lung, specifically focused on lineage markers. That is, we wanted to distinguish between heterogeneity that might be induced by the environment (e.g. inflammation or hypoxia) and focus only on heterogeneity derived from promiscuity amongst transcription factors specifying canonical epithelial cell types in the lung. Thus we define *Lineage Phenotypic Volume,* by adapting a metric of phenotypic heterogeneity, Phenotypic Volume[13] (described below) to compare the extent of Phenotypic Volume occupied by lineage specific genes in epithelium derived from the normal lung and primary tumors, within a limited lineage related gene-set.

More specifically, the metric of Phenotypic Volume as defined in [13] is the pseudo-determinant of the gene-gene covariance matrix. Thus, this metric for volume considers covariance between all gene pairs, in addition to their variance, to measure the volume spanned by independent phenotypes. To describe this metric intuitively, consider the case with only two phenotypes: the determinant is equal to the area of the parallelogram spanned by two vectors representing the phenotypes (in our case defined by expression of lineage specific genes). This area is larger for independent phenotypes, but is equal to zero if they are fully dependent. With more than two phenotypes, we are then interested in measuring the volume of the parallelepiped spanned by all these phenotypes, where more independent covariance patterns lead to increased volume. The volume of the parallelepiped spanned by these phenotypes is measured by the pseudodeterminant of the covariance matrix, which can be computed as the product of its nonzero eigenvalues.

The Phenotypic Volume (number of observed cell-states) is naturally correlated with the size of a population. Therefore, to correct for the effect of cell number differences across groups when comparing their volume, the same number of cells was uniformly sampled with replacement from each group in the comparison. The gene-gene covariance matrix was then empirically computed for each randomly selected subset of cells. Imputed data was not utilized here because it could alter the gene-gene covariate structure. The log Phenotypic volume was then computed as the sum of the log of the non-zero eigenvalues, λ_e, of each empirical gene-gene covariance matrix:

$$\log(Phenotypic\ Volume) = \sum_{e} 0.5*\log_{10}\left(\lambda_e^2\right);\ \forall\ \lambda_e > 0\,.$$

Finally, given the high number of dimensions (genes), the log of the Phenotypic Volume was normalized by the total number of genes.

Rather than computing Phenotypic volume on all genes, Lineage Phenotypic Volume restricts the computation to lineage-specific genes, including all genes annotated in Supplementary Table 1 that are variably expressed within the merged data (n = 833). We computed the *Lineage Phenotypic Volume* per data subset (primary verses tumor) as described above, sampling 500 cells, repeated 50 times to achieve the range of values reported in (Fig. 2g).

Diffusion component analysis and extrema: Diffusion maps were used to characterize major components of variation across cells[70] and the extrema of the top three most informative diffusion components were determined as bounding states of the phenotypic space. As described in[14], a cell-cell Euclidean distance matrix was computed based on the principal components of the normalized (unimputed) count matrix, where the number of principal components was selected based on the knee point of the cumulative explained variance[49]. An adaptive Gaussian kernel was then applied to convert distances into affinities, so that similarities between two cells decreases exponentially with distance. The affinity matrix was then row-normalized to construct a Markov transition matrix, whose eigenvectors are termed diffusion components. The eigenvalues of this matrix provide information on the importance of each diffusion component. In this dataset, we focused on the top three diffusion components based on the Eigen gap of the ranked diffusion components.

Next, we used the top three diffusion components to define bounds on the phenotypic space, where each component represents dominant axes of variation and its extreme ends define boundaries of the observed phenotypic states. To determine which cells belonged to the extreme (bounding) states determined by the ends of each diffusion component, we used the maximum of the second derivative of cells ranked along each diffusion component. For each of the two ends of the diffusion component, we defined all cells beyond this maxima (above for the top half and below for the bottom half) to constitute the extreme ends (Extended Data Fig. 7d). Next, the distribution of cell source (normal vs. tumor) and lineage was evaluated within the extreme end of each diffusion component (Extended Data Fig. 7e–f). Gene trends along each diffusion component were identified by evaluating the Pearson correlation between gene expression and the ordering of cells along each diffusion component (Supplementary Table 7). Gene expression trends were computed using imputed data to prevent dropout from adversely affecting the trends and were smoothed using a 20-cell sliding window. Finally, GSEA was performed on genes ranked by their correlation with each diffusion component, and gene signatures associated with each diffusion component ($p_{adj} < 0.25$) are listed in Supplementary Table 9.

Metastatic clusters ranked by lung epithelial development: Of the 21 Phenograph clusters identified in epithelium from the normal lung, primary tumors and metastases (analysis of subset *E3*, described above), 11 of these clusters were sourced >10% from all cells derived from the 5 metastatic samples and were termed patient metastatic clusters (Extended Data Fig. 8a, indicated with a star). These patient metastatic clusters differentially expressed gene signatures related to migratory stem cells, respiratory system development and morphogenesis, and downstream regenerative alveolar epithelial progenitors (Extended Data Fig. 8a). This informed selection of gene-sets (listed in Supplementary Table 2) used to characterize these clusters (Fig 3a). First, the 11 patient metastatic clusters were ranked by their average expression of a 34-gene signature of lung epithelial development computed on the imputed and normalized count matrix. Boxplots show the distribution of this score across cells within each patient metastatic cluster; clusters are ranked in ascending order from left to right by the distribution median (Fig 3a). Then, the average expression of other key gene signatures, again computed on the imputed and normalized count matrix, were visualized using boxplots per patient metastatic cluster in the same ranked order. To test whether the ranking of metastatic clusters was robust to individual genes in the 34-gene signature of lung epithelial development, we performed a leave-one-out analysis, whereby we iteratively removed each gene from the GO lung epithelial development signature and ranked metastatic clusters as described above using the remaining 33 genes (Extended Data Fig. 8f). Of the 34 gene analyzed, 19 had no change to order and 10 genes showed a swap of adjacent clusters in the ordering. We observed more frequent swapping between Clusters H13 and H7 or between Clusters H6 and H10, as they had similar distributions. The only outlier to this was gene *AGR2*, the highest expressed gene in this signature. Removal of *AGR2* significantly altered the ranking of two clusters: H13 and H14. However, expression of *AGR2* alone was not sufficient to drive metastatic cluster ranking, which indeed reflects the overall average expression of the cumulative lung epithelial development signature. Based on these genesets we observed patient metastatic clusters partitioned into 3 metastatic classes (type I-III), driven by their expression of adult stem, lung morphogenesis and alveolar epithelial progenitor programs. Finally, a Mann-Whitney U test was computed per meta-class for each expression signature by pooling cells from all patient metastatic clusters assigned to each meta-class and comparing to all remaining cells. As always, reported p-values are computed on the normalized (un-imputed) count matrix.

Assigning individual tumor cells to developmental states: Individual cells were likewise ranked in ascending order based on their cumulative expression of the same lung epithelium development signature computed on the imputed and normalized count matrix. Expression of key endoderm- and lung-specifying transcription factors along ranked cells are visualized on a heatmap (Fig. 3c); where individual gene expression levels were z-normalized across cells and smoothed using a 20-cell moving window. Additionally, we report the average expression of three canonical proliferation markers: proliferating cell nuclear antigen (*PCNA*), *MKI67* and minichromosome maintenance complex component 2 (*MCM2*) below the ranked tumor cells. Phenograph clustering (k = 500) was applied directly on this matrix. Six unique clusters were identified and assigned to the type I proliferating (I-P) or quiescent (I-Q) developmental state (distinguished by proliferation score), 2 of these clusters were collapsed and assigned to the type II developmental state, and the final 2 clusters were

collapsed and assigned to type III developmental states. Clusters were collapsed based on similar gene expression of key endoderm- and lung-specifying transcription factors (Fig. 3c) to simplify our description of developmental states observe in patient tumor cells.

Computing overall survival hazard ratios: To identify patient metastatic clusters whose abundance in primary tumors portends a poor prognosis (Fig. 3f), the mean expression of the top 10 DEGs (all probes/gene) per cluster was used to analyze overall survival (OS) in the lowest (Q1) vs. highest (Q4) patient quartiles using Kaplan-Meier Plotter[37]. DEGs identified by MAST were ranked according to: $rank = -10 * log_{10}(p_{adj}) * sgn[log_2(fold\ change)]$ and genes related to translation and ribosomal RNA transcription and processing (listed in Supplementary Table 2) were removed from the ranked genes. The OS hazard ratio (HR) is reported with confidence intervals per metastatic cluster. HR > 1 are poorly prognostic; HR < 1 indicated improved OS and any CI crossing the line at 1 are not significant.

Data analysis of scRNA-seq from xenograft model of metastasis: An advantage to SEQC is that it is able to process 10X, in-drops and drop-seq using the same pipeline. The same SEQC strategy for filtering high quality cells, constructing count matrices from reads and selecting abundant genes (described above under *Pre-processing, cell selection and filtering of droplet-based scRNA-seq data*) was applied to our xenograft mouse model of human metastases. The only exception being that when filtering viable cells from ambient mRNA, we adjusted the width of the step size taken to compute the second derivative of the empirical cumulative density function of total cell transcript counts because the shape of this cumulative function could be bimodal for data acquired using inDrop. This yielded a total of 8,748 tumor cells with a median library size of 3,423 transcripts per cell from spontaneous metastatic derivatives sequenced after one passage of *in vitro* antibiotic selection (essential to isolate putative DTCs, Fig. 4 Data) and 6,073 tumor cells with a median library size of 3,399 transcripts per cell sequenced immediately upon dissociation from NK cell-depleted metastases (Fig. 6 Data). These two datasets were processed separately as described below.

Each count matrix was normalized for library size per cell, whereby the expression level of each gene was divided by the cell's total library size and then scaled by the median library size of all cells. Principal Component Analysis (PCA) was then computed using randomized principal component analysis[48] applied to the normalized count matrix. Next, ordinary least squares was applied to linear regress library size out of each principal component because a partial correlation was observed between some principal components and library size. Finally, MAGIC imputation[14] was applied to the median-normalized count matrix to further denoise and recover missing gene values using conservative parameters ($t = 3$, $k = 27$). Imputation was performed using the top principal components of the normalized count matrix, selected based on the knee point of the cumulative explained variance[49]. The number of selected principal components were 28 and 34 in the spontaneous and NK cell-depleted metastases respectively; explaining ~60% of variance in each dataset. Principal components were then re-computed on the imputed data and 18 and 32 principal components were selected in the spontaneous and NK cell-depleted metastases respectively; explaining > 90% of the data variance. After imputation, a correlation between some principal components and library size was no longer observed.

Force-directed graphs[51] for visualization and Phenograph clustering using a k-NN graph[15] were computed on the selected principal components of the imputed data. The number of nearest neighbors, *k*, used to construct the k-NN graph for Phenograph was selected such that the Jaccard graph per data subset was fully connected ($k = 7$ for both datasets) and using a minimum value of *k* for which cluster assignments were stable, as measured by the adjusted rand index (ARI)[15] for categorical cluster assignments across pairwise values of *k* for all cells. For both datasets, 18 clusters were identified using k = 70 (Fig. 4d and Extended Data Fig. 10d).

Finally, DEGs per Phenograph cluster compared to all other clusters were identified using the R package MAST[44] and ranked as described above for GSEA. DEG per Phenograph cluster from spontaneous metastatic derivatives, ranked according to their Bonferroni-corrected *p* value and fold change, are annotated in Supplementary Table 12; associated differentially expressed pathways distinguished by $p_{adj} < 0.25$ are also provided in Supplementary Table 13. Likewise, DEG per Phenograph cluster from NK cell-depleted metastases, ranked according to their Bonferroni-corrected *p* value and fold change, are annotated in Supplementary Table 14; associated differentially expressed pathways distinguished by $p_{adj} < 0.25$ are also provided in Supplementary Table 15.

Assigning mouse metastatic clusters to human developmental states: Variably expressed genes were identified within each mouse and human dataset separately based on the dispersion of each gene across all cells as described above. 2096 intersecting genes were variably expressed in spontaneous mouse metastatic clusters and in human patient data. Likewise, 2895 intersecting genes were variably expressed in NK cell-depleted mouse metastatic clusters and in human patient data. Intersecting genes were then used to compute the correlation between the genome-wide expression patterns of mouse Phenograph clusters with the four development stages observed in human tumors (annotated in Fig. 3c as type I-Q, I-P, II, III). First, we computed the median expression of each gene per developmental state on the patient normalized count matrix, including all cells, from all clusters assigned to that developmental state in our computation. Then we centered the expression of each gene across all developmental states by mean subtraction. Ultimately, this provided a centered reference for each of the four developmental stages observed in patient tumors onto which we would like to map mouse Phenograph clusters identified in spontaneous and NK cell-depleted metastasis respectively. For each mouse dataset, we similarly computed the median expression of each gene per mouse Phenograph cluster on its normalized count matrix and centered the expression of each gene across mouse Phenograph clusters by mean subtraction. Finally, we computed the Pearson correlation between the centered transcriptional profile of each mouse Phenograph cluster and each patient developmental state in a pairwise manner. For each pairwise correlation, the Python package scipy was utilized to compute a *p* value testing for non-correlation. We visualized the genome-wide correlation between these two independent sets using a bipartite graph (see Customized plotting functions), whereby edges link significantly correlated mouse Phenograph clusters (circular nodes) and human developmental states (square nodes). Correlations were considered significant for Pearson $R > 0.20$ and $p < 0.05$; edge width is scaled by the magnitude of the correlation. The same criteria were applied to assign mouse metastatic

clusters to patient developmental states when reporting the fraction of cells assigned to each developmental stage per spontaneous and NK cell-depleted mouse macrometastases (Fig. 6c). Clusters not meeting these criteria were simply marked as un-assigned. When visualizing the mapping between spontaneous metastases and developmental states observed in patient tumors, the circular nodes representing mouse Phenograph clusters are shown as pie charts representing fraction of cells per mouse sourced from DTCs, incipient and macro-metastases (Fig. 4f).

Analysis of additional validation datasets—The E8.75 gut tube pseudo-space ordering and imputed gene expression data was downloaded from[45] (https://endoderm-explorer.com). Gene expression trends of the selected genes were estimated using Generalized Additive Models by fitting the imputed expression as a function of the pseudo-space ordering[71]. The estimated trends were z-transformed across the pseudo-space to be used for plotting.

Lung adenocarcinoma TCGA normalized data was downloaded from cBioPortal (https://github.com/cBioPortal/datahub/blob/master/public/luad_tcga/data_RNA_Seq_v2_expression_median.txt). The Spearman Rank correlation was used to evaluate the strength and directionality of the relationship between SOX9 target genes (Supplementary Table 2) and the average expression of MHC Class I genes (Supplementary Table 2) (Fig. 5j). The $\log_2$ average expression of an NK cell-specific signature[72] (Supplementary Table 2) was also evaluated in the first and third tertile of patients stratified by their SOX2 or SOX9 expression (Extended Data Fig. 9a).

Normalized mRNA data from all mouse KP derivatives was downloaded from [39] and *Sox2* and *Sox9* expression levels per sample were plotted directly without additional data pre-processing (Extended Data Fig. 9c).

Cell lines—H2087 (ATCC) latency competent cancer cell (LCC) derivatives (H20870-LCC1 and H2087-LCC2[16]) were cultured in RPMI 1640 media supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, 100 IU/ml penicillin/streptomycin, 1 μg/ml amphotericin B, 0.5 mM sodium pyruvate, 10 mM HEPES, 50 nM hydrocortisone, 25 nM sodium selenite, 20 μg/ml insulin, 10 μg/ml transferrin, 0.5% bovine serum albumin (BSA), and 1ng/ml recombinant human epidermal growth factor. LCC derivatives were re-inoculated in mice to confirm their latent phenotype. KP482T1 metastatic cells (courtesy of Tyler Jacks' lab) were cultured in RPMI 1640 media supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, 100 IU/ml penicillin/streptomycin and 1 μg/ml amphotericin B. No commonly misidentified cell lines were used and all cell lines tested negative for mycoplasma contamination.

Overexpression constructs—FUW-tetO-h*SOX2* (Addgene plasmid #20724), was subcloned into the pLVX-tight-puro lentiviral expression vector using EcoRI restriction enzyme site. *SOX9* was amplified from the human *SOX9* ORF clone (GenScipt, Cat. OHu19789D) and cloned into the pLVX-tight-puro vector using NotI and XbaI restriction enzyme sites.

Immunoblotting—Cells in described conditions were lysed in RIPA cell lysis buffer (Cell Signaling Technology) containing protease inhibitors (Roche, cOmplete, mini, EDTAfree protease inhibitor tablets, Cat. 11836170001) and phosphatase inhibitors (Thermo Scientific, Halt Phosphatase Inhibitor Cocktail, Cat. 78427). Protein concentrations were determined with the BCA Protein Assay (Pierce). Proteins were separated in NuPAGE Novex 4–12% Bis-Tris gels using 1X MOPS SDS running buffer, and transferred to nitrocellulose membranes. Membranes were immunoblotted with primary antibodies against SOX2 (Abcam, Cat. ab97959), SOX9 (Abcam, Cat. ab185966) and β-actin (Cell signaling Technology, Cat. 3700). Proteins were detected using IRDye secondary antibodies captured on an Odyssey CLx infrared imaging system (LI-COR Biosciences).

Reverse-transcriptase quantitative polymerase chain reaction (RT-qPCR)—Total RNA was isolated using the RNeasy Plus kit (Qiagen) and eluted in 60 ml H_2O. RNA concentrations were estimated using a NanoDrop analyzer (ThermoFisher). 1 mg purified RNA was reverse transcribed using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Total cDNA was diluted 1 in 10 with nuclease free H_2O. 2 ul of the diluted cDNA was used in a 10 ul qPCR reaction using the Total TaqMan® Universal PCR Master Mix (2X) (Applied Biosystems) and TaqMan™ Gene Expression Assay (FAM) (20X) (Applied Biosystems). Quantitative PCR was performed on ViiA 7 Real-Time PCR System (Life Technologies). Assays were used for *SOX2* (Hs01053049-s1), *SOX9* (Hs01001343_g1), *HLA-A* (Hs01058806_g1), *HLA-B* (Hs00818803_g1), *B2M* (Hs00187842_m1), *ULBP1* (Hs04194603_s1), *ULBP2* (Hs00607609_mH), *ULBP3* (Hs00225909_m1), *RAET1E* (Hs01026643_g1), *RAET1G* (Hs01584111_mH), *RAET1L* (Hs00867544_gH), *MICA* (Hs00741286_m1) and *MICB* (H200792952_m1). The ddCt method was used to calculate relative expression values, which were normalized to the housekeeping gene *ACTB* (Hs01060665_g1) or *GAPDH* (Hs02786624_g1).

Cell surface protein expression analysis—Adherent cells were detached using 0.25% Trypsin to generate single cell suspensions, and re-supended in FACS buffer (1X PBS, 0.25mM EDTA, 2% FBS). Cells were incubated with APC-conjugated antibody for HLA-Class I Bw4 for 20–30min (Miltenyi Biotec, Cat. 130–103-918), and washed twice in FACS buffer. Cell surface expression of HLA-Bw4 was analyzed by flow cytometry on a BD Fortessa (BD Biosciences). DAPI (Thermo Fisher Scientific, Cat. D1306) was used to exclude dead cells and IgG control staining was used as a negative control to set up the gate for analysis.

RGB labeling to track clonal dynamics of metastases—To measure the size and number of subclones present in micrometastic lesions over time in xenograft models, we stochastically expressed trichromic reporter vectors in single cells using three lentiviral gene ontology vectors encoding Cerulean, Venus and mCherry fluorescent proteins at 50–60% multiplicity of infection[43]. Transduced cells were individually marked by an extensive color palette determined by variations in vector copy number and insertion sites, shown to be stable after multiple passages in vitro and in vivo[43]. Cerulean, Venus and mCherry fluorescent proteins in RGB-labeled cells were detected based on their intrinsic fluorescence.

Animal studies—Animal experiments were performed in accordance with protocols approved by the MSKCC Institutional Animal Care and Use Committee. 1×10^5 H2087-LCC2 cells[16] expressing a lentiviral TK-green fluorescent protein (GFP)-luciferase (TGL) construct and the antibiotic resistance marker Blasticidin in 100 μl of PBS, were injected into the left cardiac ventricle of 6–7 week-old female athymic NCR nu/nu mice (Charles River) to ensure systemic dissemination of tumor cells. Metastatic colonization was then measured weekly using bioluminescence imaging (BLI). Mice were injected retro-orbitally with D-luciferin (150 mg/kg), anesthetized with isofluorane and subjected to BLI using an IVIS Spectrum Xenogen instrument (Caliper Life Sciences). BLI images were analyzed using Living Image Software v.4.4. NK cell depleted-outbreaks were achieved by twice-weekly intra-peritoneal injection of 100μl of anti-asialo-GM1 antibody (Wako Chemicals Cat. 98610001) from the start of the NK cell depletion regimen. NK cell depletion was monitored by analysis of Lineage-CD45+Nkp46+ (Lineage-: CD3-Tcrβ-CD19-B220-CD11c-Ly6G-F4/80-) NK cells from peripheral blood using flow cytometry (BD Fortressa, BD Biosciences). Experiments were not blinded or randomized.

Cancer cells were freshly derived from DTCs, incipient and spontaneous metastatic outbreak *in vivo* immediately following sacrifice of animals +/− BLI-detected metastases. Ex-vivo BLI was performed on harvested organs to confirm the absence or to define the precise location of macrometastatic lesions. Organs were resected under sterile conditions and mechanically dissociated using a gentleMACS dissociator (Miltenyl Biotec) and placed in culture medium containing a 1:1 mixture of DMEM/Ham's F12 medium supplemented with 0.125% collagenase III and 0.1% hyaluronidase. Minced samples were incubated at 37°C for 1 h, with gentle rocking to produce single cell suspensions. After collagenase treatment, cells were briefly centrifuged, re-suspended in 0.25% trypsin, and incubated for a further 15 min at 37°C. Cells were then re-suspended in their culture conditions and allowed to grow to confluence on a 15-cm dish in selection media containing blasticidin to select for tumor cells and exclude host cells. scRNA-seq assays and immunofluorescence assays were performed after a single passage from the primary dissociation.

Single cell suspensions were similarly derived from metastatic outbreaks in NK cell-depleted mice, but were processed immediately for scRNA-seq without antibiotic selection. Cell suspensions were subsequently flow sorted with a BD FACSAria II cell sorter fitted with a 100 μm nozzle to enrich for viable, GFP-positive single cells according to forward and side scattering, fluorescent protein expression, and DAPI exclusion. Cells were sorted directly into RPMI media with 10% FBS, washed thrice and re-suspended in PBS containing 0.04% BSA for single cell encapsulation.

NK cell cytotoxicity assays—Splenocyte suspensions were prepared by mechanical dissociation. NK cells were purified by magnetic depletion of non-NK cells using NK Isolation Kit II (Miltenyi Biotec, Cat. 130–096-892) and separation with magnetic columns (Miltenyi Biotec, Cat. 130–042-401). NK cells were cultured overnight in NK cell media (RPMI 1640 medium supplemented with 10% FBS, β2-mercaptoethanol, non-essential amino acids, 10 mM HEPES, 0.5 mM sodium pyruvate, 2 mM L-glutamine, and 10 IU/ml penicillin/streptomycin) containing 1000 U/ml recombinant interleukin-2.

For measuring NK killing of cancer cells, target GFP-expressing cancer cells were incubated with or without NK cells at a 1:10 carcinoma:NK cell ratio, for 3 h at 37°C. Cell mixtures were stained with 7-AAD Viability Staining Solution (BioLegend, Cat. 420404) and Annexin V (BioLegend, Cat. 640941) to assess tumor cell cytotoxicity by flow cytometry; NK cells were excluded from the analysis by immunostaining with anti-CD45 antibody (BD Biosciences, Cat. 565967). Orthogonally, the relative proportion of SOX2-expressing and SOX9-expressing cancer cells after NK killing was quantified by image cytometry analysis of SOX2 and SOX9 immunofluorescence (described below). Experimentally, cancer cells were plated in cell chamber slides overnight, and incubated with or without NK cells (effector to target ratio of 1:10) for 21 h at 37°C. Slides were washed with PBS, fixed in 4% PFA for 10 min at room temperature, then washed with PBS thrice, followed by immunostaining for SOX2 or SOX9. Slides were imaged with a Leica TCS SP5-II inverted point-scanning confocal microscope with a 20x/0.7NA objective with 1.5x optical scan zoom. Between 5 and 8 locations per sample were imaged and quantified.

Immunofluorescence—Harvested organs were fixed in 4% paraformaldehyde (PFA) overnight at 4°C and washed in PBS. Organs were cryoprotected by immersion in 15% then 30% sucrose. Cryoprotected organs were mounted using OCT Compound (Sakura) onto a sliding microtome outfitted with a platform freezing unit (Thermo Scientific, Microm KS-34 and Microm HM-450). 80μm sections were cut and sequentially stored in anti-freezing solution (30% ethylene glycol, 30% glycerol in PBS) at −20°C. For macrometastatic lesions, PFA-fixed tissues were washed 2–3 times in PBS, dehydrated in 70% ethanol, paraffin embedded and prepared as 5μm sections. Macrometastases in the bone were fixed with 4% PFA overnight, washed in PBS and incubated with EDTA-based decalcification solution (140 g/L EDTA, pH 7.4) for 1–2 weeks at 4°C with agitation and daily changes of decalcification solution. Tissues were then washed with water for 1 hour, re-fixed with 4% PFA, washed twice in PBS, followed by ethanol dehydration, paraffin embedding and sectioning. Immunofluorescence staining for SOX2 and SOX9 was performed at the Molecular Cytology Core Facility, MSKCC, using the Discovery Ultra processor (Ventana Medical Systems-Roche)[73]. A rabbit polyclonal anti-SOX2 antibody (Abcam cat#97959) was used at 5 μg/mL concentration. A rabbit monoclonal anti-SOX9 antibody (Abcam cat#185966) was used at 1 μg/mL concentration. Samples were incubated with the primary antibodies for 5 h, followed by a 60-min incubation with biotinylated goat anti-rabbit IgG (Vector labs, cat#:PK6101) at a concentration of 5.75 μg/mL. Detection was achieved by Streptavidin-HRP D (Ventana Medical Systems), followed by incubation with Tyramide-Alexa Fluor 568 (Invitrogen, cat. #B40956). Specificity of antibodies during co-immunofluorescence were validated using a panel of cell lines validated to be double negative, double positive, or single positive for SOX2 and SOX9.

Image Cytometry Analysis—For all *in vitro* assays, the centroids of DAPI-stained nuclei were identified using a MATLAB implementation of the IDL tracking methods developed by John Crocker, David Grier, and Eric Weeks (physics.georgetown.edu/matlab/). Nuclei were then segmented using intensity thresholding of DAPI staining followed by a watershed process combining DAPI intensity with centroids determined in the previous step. Segmented nuclei were filtered by size using a threshold set using Otsu's method[74] and

circularity to exclude NK cells and debris. Regions with poor segmentation due to debris or with exclusively NK cell populations were manually excluded from further analysis. For quantitation of all *in vivo* tissue immunofluorescence, a convolution neural network based on the instance based Mask-RCNN architecture[75] implemented in Tensorflow was used for automated nuclear segmentation. Segmented nuclei were then used as masks for per-nucleus quantification of the cumulative intensity of SOX2 and SOX9 normalized to DAPI intensity. For SOX2 and SOX9 co-immunofluorescence experiments, digital compensation was performed to correct for incomplete separation of fluorophores using a combination of single-stained controls and double-stained control populations with the assumption that control populations possess a non-zero fraction of cells single-positive for each marker[76]. Thresholds for positive and negative staining of each antigen were set based on the mean and standard deviation of each fluorophore's distribution. Analysis and plotting were performed in Python.

Statistics—All statistical tests are explicitly and comprehensively described in their corresponding figure legends. Normality was validated for all t-tests; otherwise, non-parametric Mann-Whitney or Kruksal-Wallis test were applied. When evaluating Pearson correlations, the Python package scipy was utilized to compute a p value testing for non-correlation. The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Data availability—Raw data from Western blots is included in Source Data. A ranked list of DEG and complete GSEA results for all Phenograph clusters analyzed in this manuscript and provided in Supplementary Tables 4–15. A custom GSEA annotation file, assembled to query cell types and pathways related lung epithelial development and regeneration is provided in the Supplementary Dataset. All raw and processed single-cell RNA sequencing data with cell type annotations was deposited in NCBI's Gene Expression Omnibus and are accessible through accession number GEO: GSE123904. Fully annotated count matrices are available for download at (https://s3.amazonaws.com/dp-lab-data-public/lung-development-cancer-progression/PATIENT_LUNG_ADENOCARCINOMA_ANNOTATED.h5 and https://s3.amazonaws.com/dp-lab-data-public/lung-development-cancer-progression/MOUSE_LUNG_ADENOCARCINOMA_METASTASIS_ANNOTATED.h5). All other datasets generated and analyzed in the current study are available from the corresponding authors upon request.

Code availability—All custom code, statistical analysis, and visualizations were performed in Python and are demonstrated in a Jupyter notebook available for download at (https://github.com/dpeerlab/lung-development-cancer-progression). The following open-source algorithms were additionally used as described in the methods: SEQC (https://github.com/ambrosejcarr/seqc), t-SNE (https://lvdmaaten.github.io/software/), MAGIC (https://github.com/dpeerlab/magic), and Phenograph (https://github.com/jacoblevine/PhenoGraph).

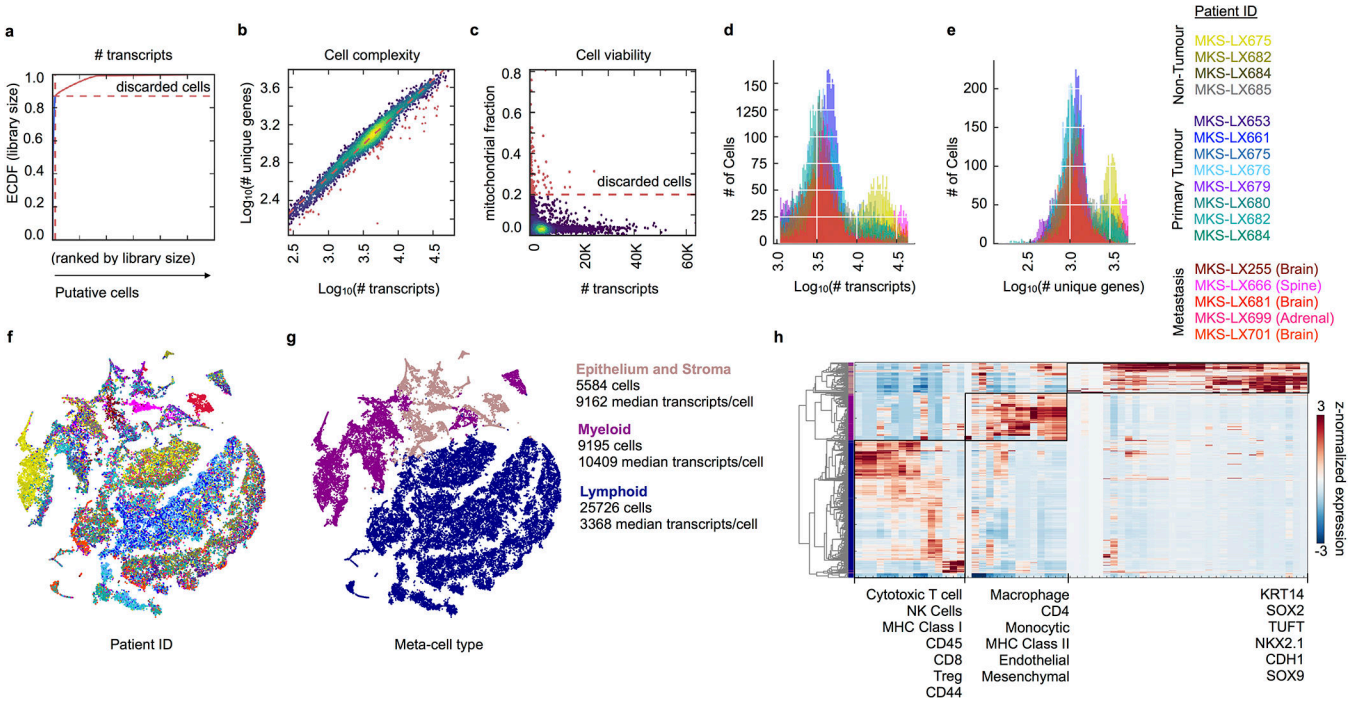
Extended Data

Patient	Tissue Site	Smoking History	Primary Tumor Size (cm)	Diagnosis	Stage	TNM Staging	Known Oncogenic Mutations	Chemotherapy
MSK-LX653	Primary	Never smoker	1.5 x 1.4 x 1.0	Lung Adenocarcinoma	IA	T1bN0M0	*EGFR, KMT2D, U2AF1, RET*	None
MSK-LX661	Primary	.05-1 ppd x 22 yr	1.3 x 0.8 x 0.7	Lung Adenocarcinoma	IA	T1bN0M0	*EGFR, TP53*	None
MSK-LX666	Bone Metastasis (T5)	Never smoker	NA	Lung Adenocarcinoma	IV	T4N3M1	*EGFR*	1. Denosumab 2. Afatinib 3.Pemetrexed + Carboplatin
MSK-LX675 **MSK-LX675**	Primary Adjacent Non-Tumor	Never smoker	4.1 x 3.0 x 2.0	Lung Adenocarcinoma	IV	T2bN1M1	*EGFR*	None
MSK-LX676	Primary	1 ppd x 40 yr (Current: 0.5 ppd)	2.1	Lung Adenocarcinoma	IA	T1bN0M0	*Negative*	None
MSK-LX255B	Brain Metastasis	1-1.5 ppd x 15-20 yr	3.4 x 2.4	Lung Adenocarcinoma	IV	T2N1M1	*EGFR, TP53*	Cisplatin + Vinorelbine
MSK-LX679	Primary	0.5 ppd x 30 yr	5.6 x 5.4 x 4.6	Lung Adenocarcinoma	IIA	T2N1M0	*KRAS*	Carboplatin + Pemetrexed
MSK-LX680	Primary	1.5-2 ppd x 35 yr	2.6 x 2.6 x 1.9	Lung Adenocarcinoma + Large Cell Neuroendocrine Carcinoma (LCNEC)	IB	T2aN0M0	*TP53*	None
MSK-LX681	Brain Metastasis	1.5 ppd x 40 yr	5.1 x 4.3 x 2.8	Lung Adenocarcinoma	IV	T2N2M1	*Negative*	Cisplatin + Vinorelbine
MSK-LX682 **MSK-LX682**	Primary Adjacent Non-Tumor	2 ppd x 40 yr	2.5 x 2.1 x 1.7	Lung Adenocarcinoma	IB	T2aN0M0	*KRAS*	None
MSK-LX684 **MSK-LX684**	Primary Adjacent Non-Tumor	0.5 ppd x 30 yr	2	Lung Adenocarcinoma	IA	T1aN0M0	*KRAS*	None
MSK-LX685	Adjacent Non-Tumor	Never Smoker	3.5 x 3.2 x 1.7	Lung Adenocarcinoma	IA	T1aN0M0	*NA*	None
MSK-LX699	Adrenal Metastasis	0.5 ppd x 10 yr	NA	Lung Adenocarcinoma	IV	T2N2M1	*KRAS, TP53*	1. Carboplatin + Taxol + Bevacizumab 2. Durvalumab + Tremelimumab 3. Nivolumab
MSK-LX701	Brain Metastasis	Never Smoker	5.0 x 3.0 x 2.2	Lung Adenocarcinoma	IV	T2N2M1	*EGFR, TP53*	Erlotinib

ppd = packs per day, yr = year

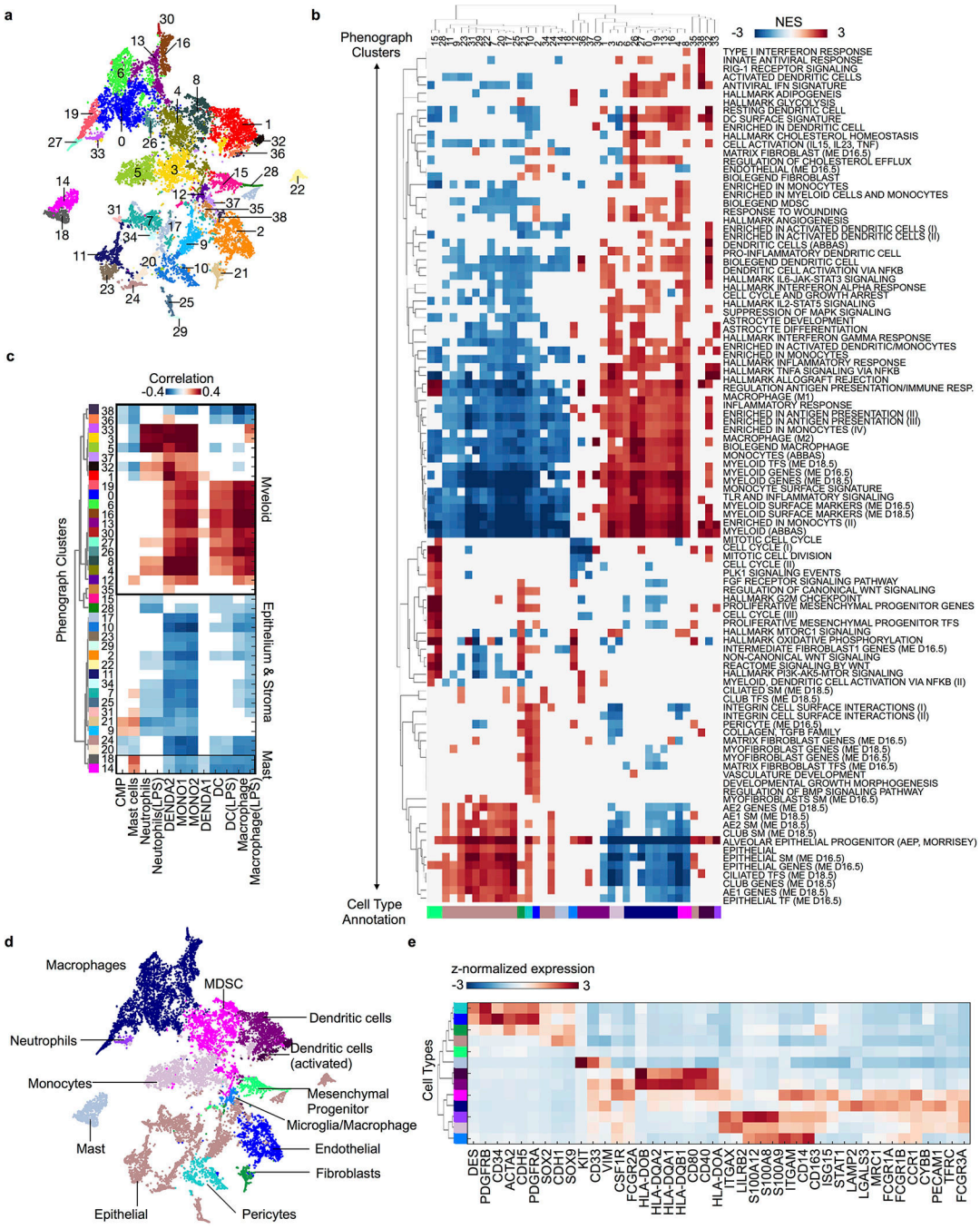
Extended Data Fig. 1: Patient attributes.

Patient resection site, smoking history, primary lesion size, disease stage, diagnostic pathology, oncogenic mutations and treatment history.



Extended Data Fig. 2: Single cell parameters and pre-processing.

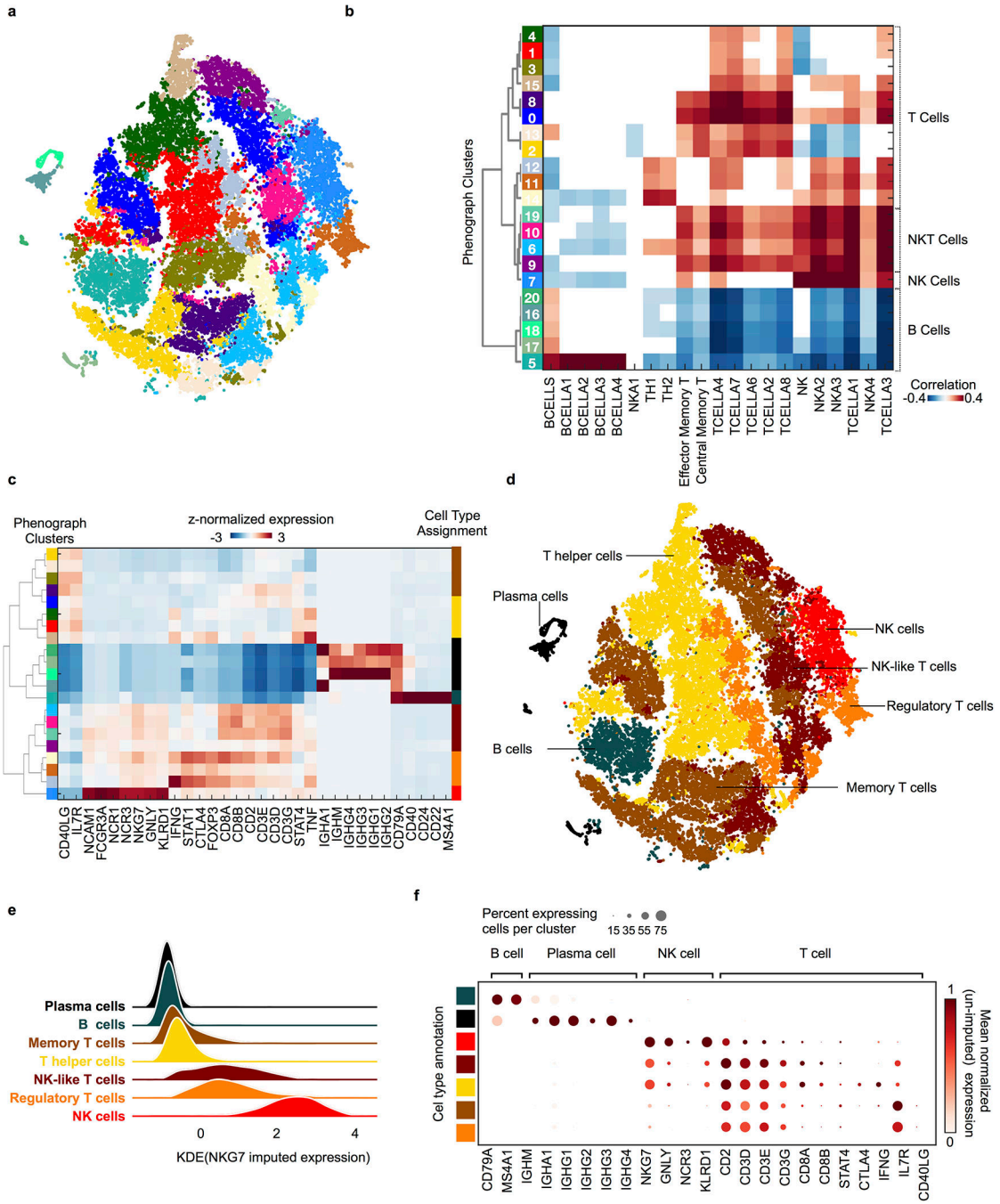
Cells were filtered based on **(a)** cumulative number of transcript counts, **(b)** cell complexity and **(c)** fraction of mitochondrial mRNA content detected per cell as described in the Methods; shown here for one representative library. Excluded cells are labeled in red. Histograms showing the distribution of **(d)** total number of transcripts detected per cell and **(e)** number of unique genes detected per cell in retained cells colored by sample. **f-h**, t-SNE projection of the complete atlas of normal lung, primary tumor and metastatic LUAD (same projection as Figure 1c, n = 40,505 cells), cells colored by **(f)** sample and **(g)** meta-cell class as determined by **(h)** unsupervised clustering of canonical gene signatures within each meta-cell class across all cells. Clustering of canonical cell type expression signatures (annotated in Supplementary Table 2), z-score normalized per gene across cells. Assignment to meta-cell classes (detailed in Methods) are colored on the dendrogram.



Extended Data Fig. 3: Phenotyping myeloid, epithelial and stromal cell types.

a, t-SNE projection of *myeloid, epithelial and stromal* cells (purple and tan populations from Extended Data Fig. 2g, n = 9,195 cells) colored and labeled by their Phenograph cluster assignment (Phenograph run on subset as detailed in Methods). **b**, Heatmap of gene signatures differentially expressed by Phenograph clusters with *abs(NES)* > 2 and p_{adj} < 0.05, clustered (rows and columns) according to the cosine distance metric for visualization; hits not meeting these criteria are whited out. Numbered by Phenograph clusters (top) and colored by inferred cell type assignments (bottom, see Methods). NES, normalized

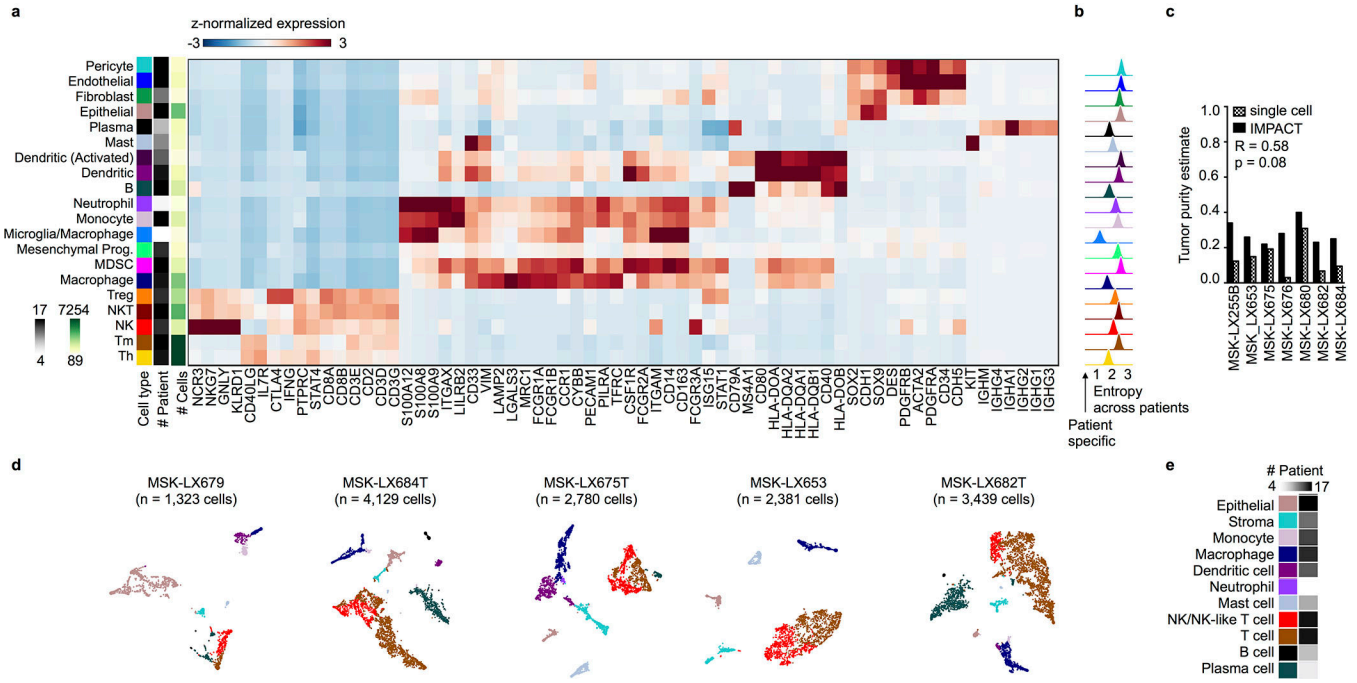
enrichment score; p_{adj}, Bonferroni corrected, two-sided p-value. **c**, Pearson correlations between Phenograph cluster centroids and bulk mRNA profiles from purified immune subpopulations[56,57] (n = 5,987 genes, Methods). Correlation coefficients are whited out if p > 0.01 for the Pearson test for non-correlation. **d**, t-SNE projection of all *myeloid/epithelial/stromal* cells (same as **a**) colored and labeled by inferred cell types. Phenograph clusters were mapped to cell types using (**b-c**) and are directly mapped back to the complete patient dataset in (Fig. 1c) using the same color scheme. **e**, Clustered heatmap of the average imputed expression per cell type of distinguishing markers, standardized by z-score. Rows are colored by annotated cell type.



Extended Data Fig. 4: Phenotyping NK, T and B cells in the lymphoid compartment.

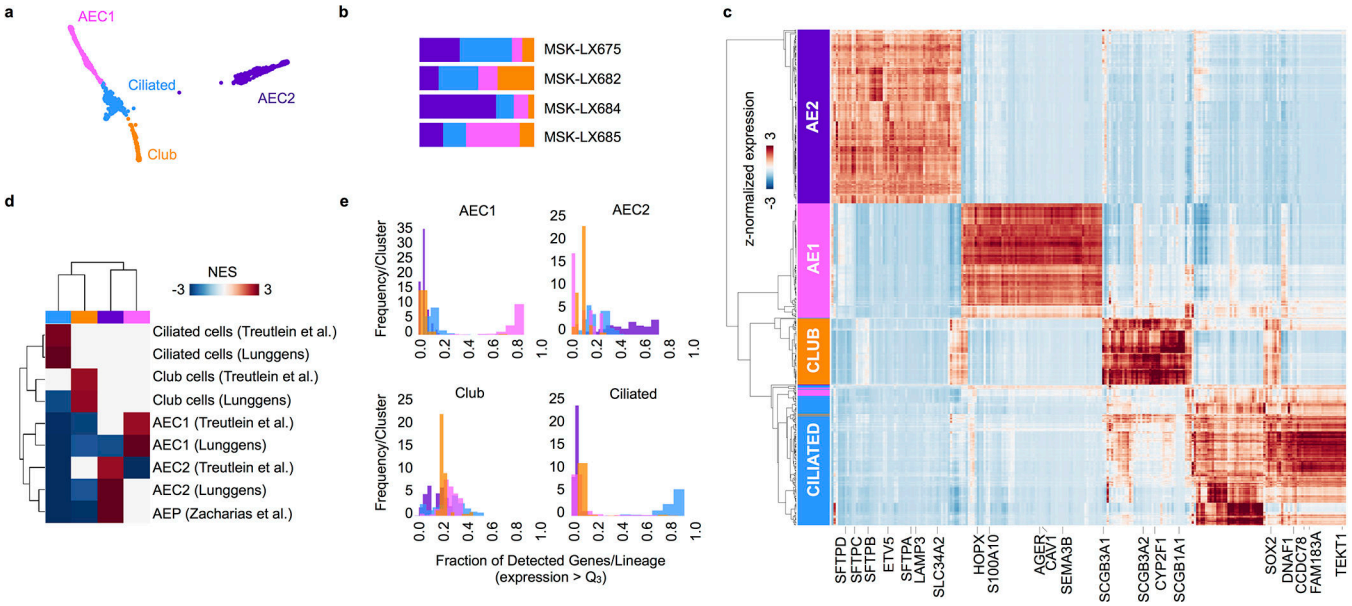
a, t-SNE projection of all *lymphoid* cells (blue cells from Extended Data Fig. 2g, n = 25,726 cells) colored by Phenograph cluster. **b**, Pearson correlations between Phenograph cluster expression centroids and bulk mRNA data published from purified immune subpopulations[56,57] computed based on intersecting, variably expressed genes (n = 5,613, Methods). Rows are colored and labeled by Phenograph clusters. Correlation coefficients are whited out if $p > 0.01$ for the Pearson test for non-correlation. **c**, Clustered heatmap of the average imputed expression per Phenograph cluster of canonical lymphoid markers,

standardized by z-scores. Rows are colored by Phenograph clusters (left) and annotated cell types (right, see Methods). **d**, t-SNE projection of all *lymphoid* cells (same as **a**) colored and labeled by inferred cell types. Phenograph clusters were mapped to cell types using (**b-c**) and are directly mapped back to the complete patient dataset in (Fig. 1c) using the same color scheme. **e**, The cell distribution of *NKG7* imputed expression, a canonical NK cell marker, across all annotated lymphoid cell types. **f,** Dot plots showing relative frequency of expressing cells and mean normalized expression (un-imputed data) of canonical markers per lymphoid cell type.



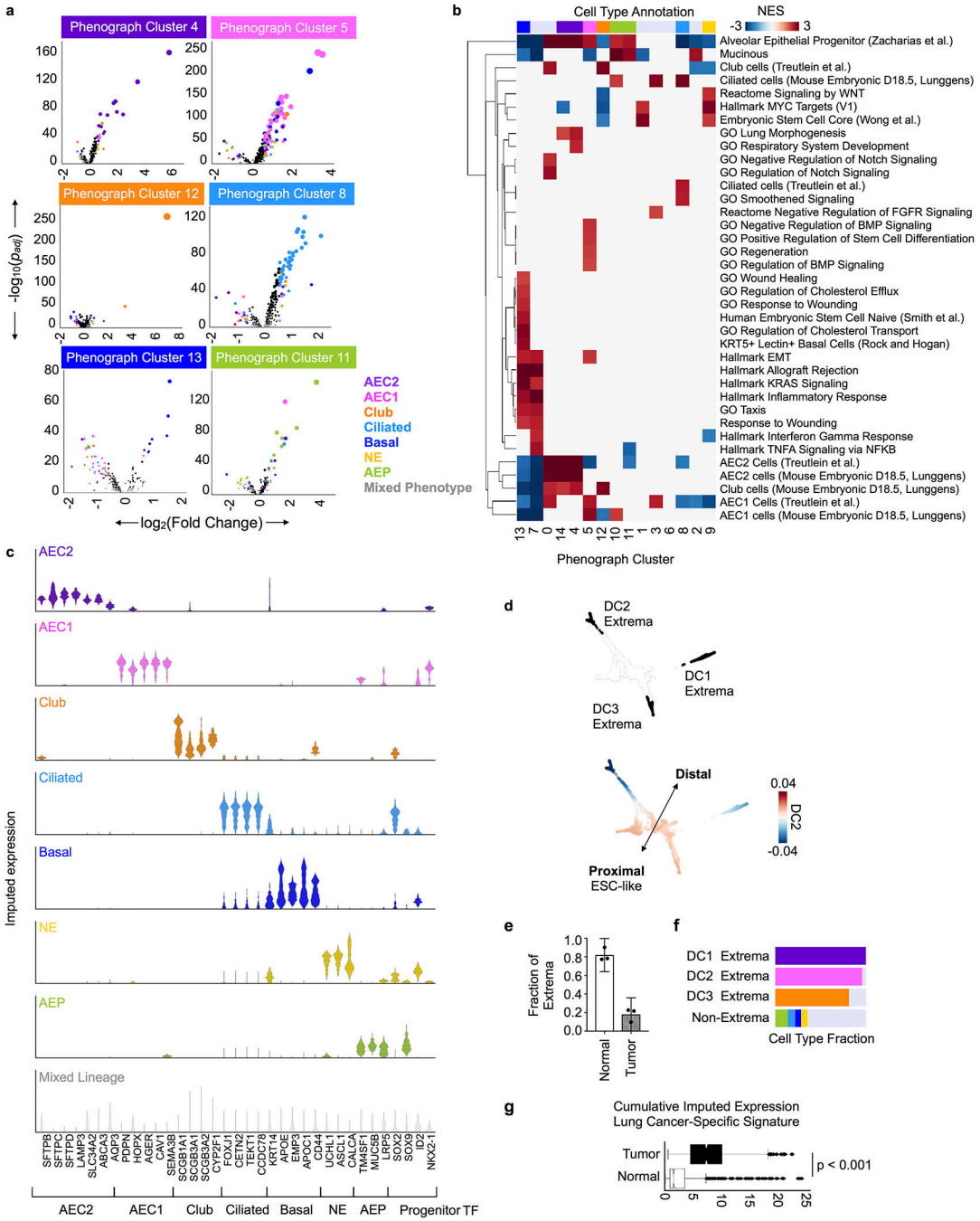
Extended Data Fig. 5: Reproducibility of cell types across patients.

a, Clustered heatmap of imputed average of key cell type markers, for all cell types annotated in the complete patient dataset (Fig. 1c) and standardized by z-scores. Rows are colored by cell type annotation, number of patients in which the cell type was detected, and the total number of cells assigned to this cell type (*left to right* on left side of clustered heatmap). **b**, Kernel density plot depicting entropy of the patient distribution as a measure of sample mixing across all patients within each cell type; computed with bootstrapping to correct for number of cells in each cluster (n = 100 random subsamples) as described in Methods. High entropy indicates most similar cells come from a well-mixed set of patient samples, whereas low entropy indicates most similar cells come from the same patient sample. Distributions are colored by annotated cell types. **c,** Estimates of tumor purity measured by scRNA-seq and targeted panel DNA sequencing of matched bulk tumor using FACETS, an allele-specific copy number analysis tool[19]. We test effectiveness of pairing between tumor purity estimates by reporting the Pearson correlation coefficient for n = 7 samples for which matched scRNA-seq and bulk, targeted panel DNA sequencing was available with one-sided P value testing for non-correlation. **d,** t-SNE projection of individual patient tumors from n = 5 representative patient samples; annotated number of cells per patient. scRNA-seq data for each tumor was processed independently as described in the Methods; each dot represents a cell colored by Phenograph clusters, labeled by inferred cell types. **e,** Number of patients in which each cell type was detected for individual patient analyses. This is concordant with patient frequencies observed in the pooled analysis, summarized in **a**. Power to detect minority cell types like neutrophils and plasma cells is reduced when analyzing patient samples one-by-one.



Extended Data Fig. 6: Epithelial cell types detected in the normal, adult human lung.

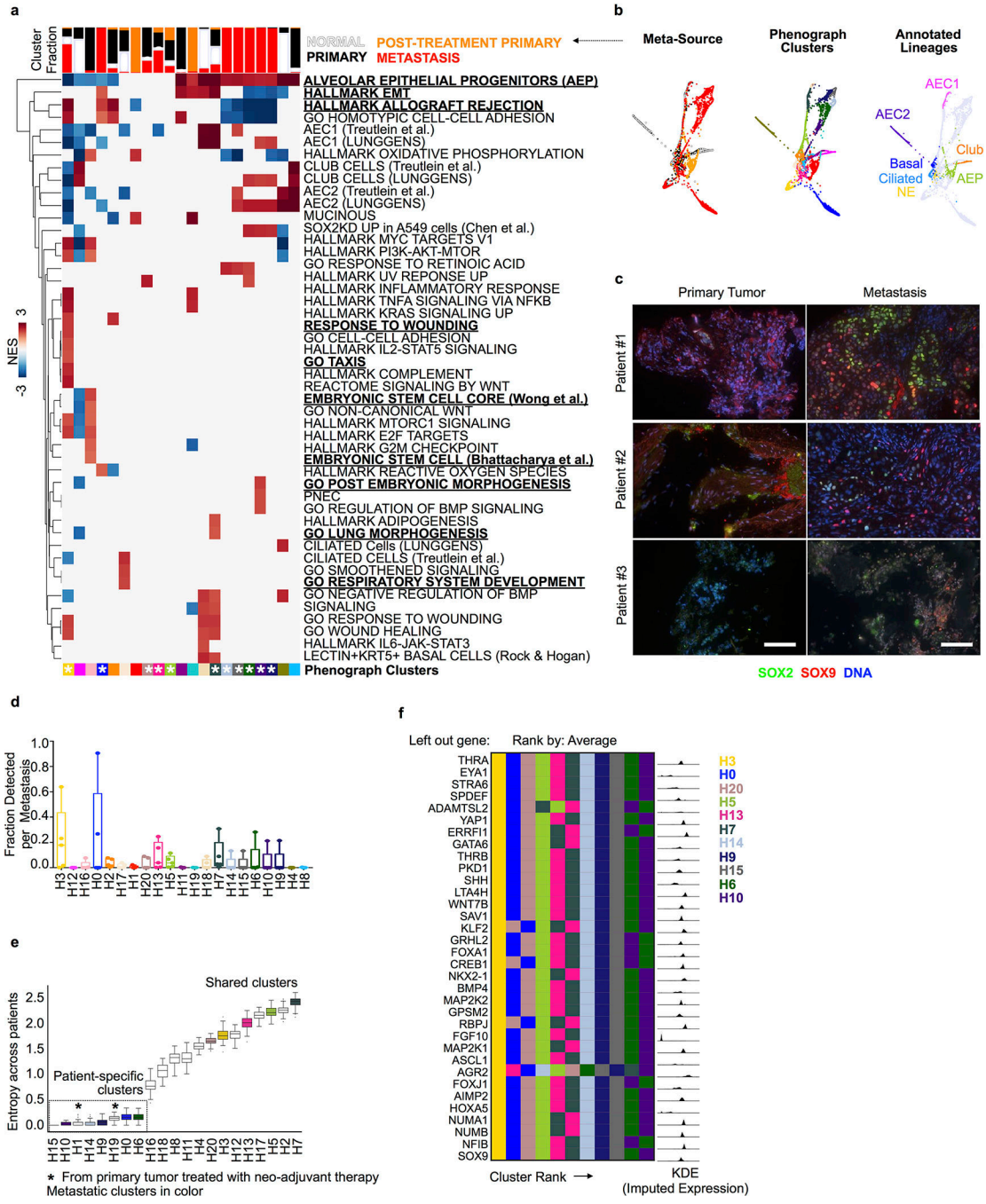
a, Force-directed layout of epithelial cells detected in the normal lung colored by Phenograph cluster and labeled with annotated cell type (n = 658 cells; Methods and below). **b**, Bar graphs representing the fraction of each cell type detected per patient. **c**, Clustered heatmap of the top 60 DEGs per cell type; imputed values standardized by z-scores are shown for visualization. DEGs were identified in un-imputed data using MAST[44] as described in the Methods. Rows are colored by Phenograph cluster and labeled by annotated cell type. **d**, A clustered heatmap of differentially expressed gene signatures within each cell type. NES is shown for pathways in which $abs(NES) > 2.5$ and $p_{adj} < 0.05$; signatures not meeting this criterium are whited out. Columns are colored by Phenograph cluster. NES, normalized enrichment score; p_{adj}, Bonferroni corrected, two-sided p-value. Complete GSEA results per Phenograph Cluster, including nominal p-values, are provided in Supplementary Table 5. **e**, Histograms showing the fraction of cells per Phenograph cluster (i.e. cell loadings) expressing at or above the 75th percentile a fraction of each cell-type specific gene signature (AEC1, AEC2, Club and Ciliated) computed on imputed data; each distribution represents cells from one Phenograph cluster. Colors associated with each annotated epithelial cell lineage are maintained as in Fig. 2.



Extended Data Fig. 7: The relationship between primary tumor and normal lung epithelial cell types.

a, Top DEGs per Phenograph Clusters annotated in Fig. 2b, compared to all other cells, computed using MAST[44]. Each lineage-specific gene is colored by associated cell type, with diameter proportional to $-log_{10}(p_{adj})$. See Supplementary Table 1 for lineage-specific genes. **b,** Clustered heatmap of gene signatures differentially enriched ($abs(NES) > 2$ and $p_{adj} < 0.05$) in one or more Phenograph Clusters. NES, Normalized Enrichment Score. Column colors (top) correspond to annotated epithelial lineages. Signatures not meeting these criteria

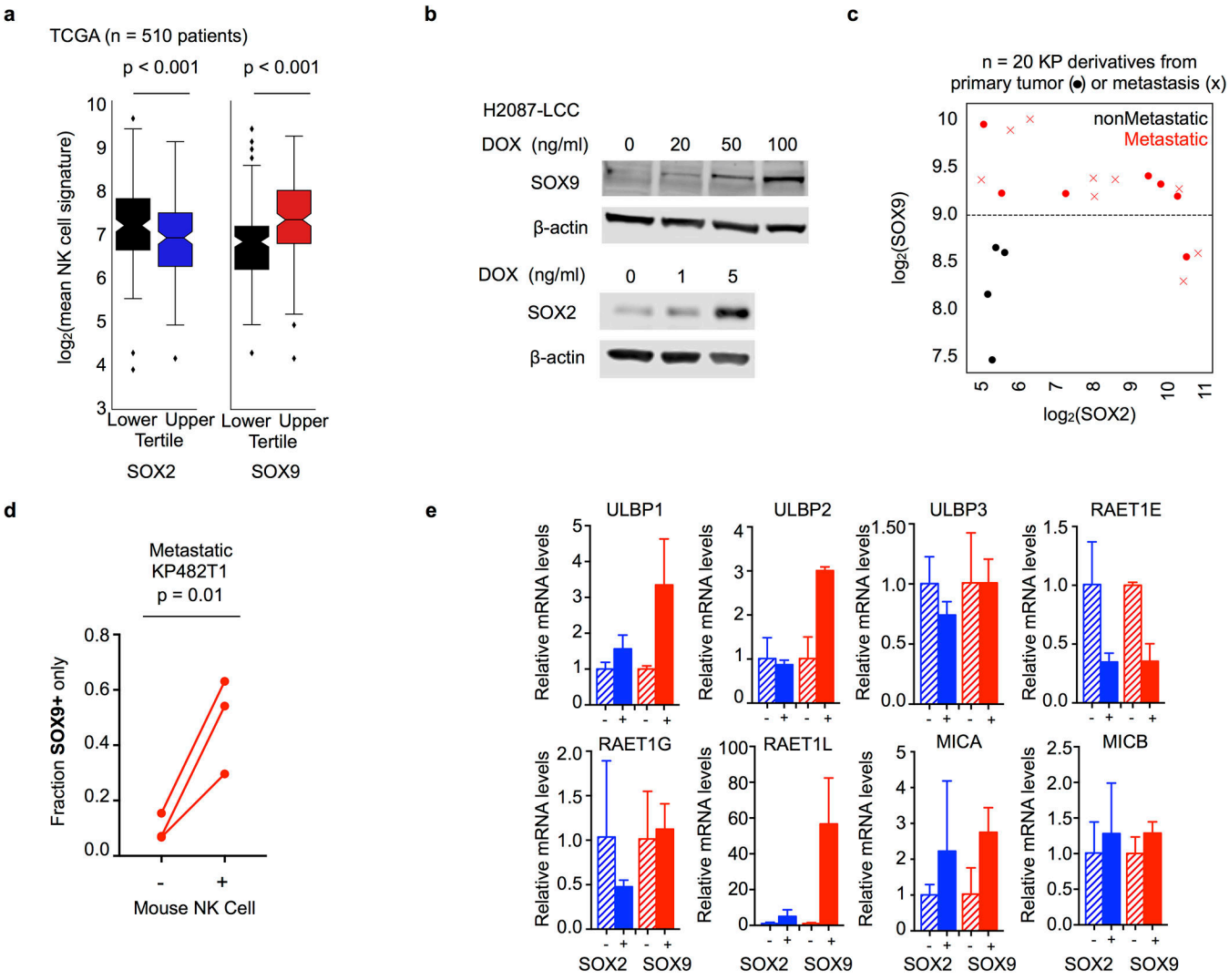
are whited out. See Supplementary Table 8 for complete GSEA results per Phenograph Cluster. **c,** Violin plots showing imputed expression of canonical lineage-specific transcription factors (columns) for each annotated epithelial cell lineage (color), scaled such that each plot has the same width; lines distinguish data quartiles. **d,** Force-directed layout of all epithelia (n = 2,140 cells) colored by extrema of the three most informative diffusion components (DCs, *above*) and by DC2 (*below*); GSEA of cells ranked along DC2 are positively enriched for embryonic stem cell gene signatures and pathways associated with proximal cell types, and negatively associated distal cell types. Complete GSEA results are provided in Supplementary Table 8. **e,** Fraction of normal- and primary tumor- derived cells comprising the union of all three DC extrema (center values, mean; error bars, 95% confidence interval; points, fraction of cells measured at n = 3 diffusion extrema). **f,** Fraction of each annotated cell type detected per diffusion component extrema and in non-extrema. **g,** Cumulative imputed expression of a bulk-derived gene signature up-regulated in LUAD and not expressed in non-cancerous epithelium[30] evaluated per cell in normal vs. tumor-derived epithelium (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, n = 2,140 individual cells); two-sided $p < 0.001$, Mann Whitney U test.



Extended Data Fig. 8: Identification of transcriptionally distinct metastatic subpopulations.

All data in this figure relates to the combined normal, primary tumor and metastatic epithelium. **a**, Clustered heatmap showing gene signatures differentially expressed across Phenograph clusters. Normalized enrichment score (NES) is colored for gene signatures in which *abs(NES)* > 1.5 and two-sided $p_{adj} < 0.05$. p_{adj}, Bonferroni corrected, two-sided p-value. Rows correspond to gene signatures, column corresponds to Phenograph clusters. See Supplementary Table 10 for complete GSEA results per cluster. Fraction of each Phenograph cluster derived per tissue source is visualized above each column. White stars

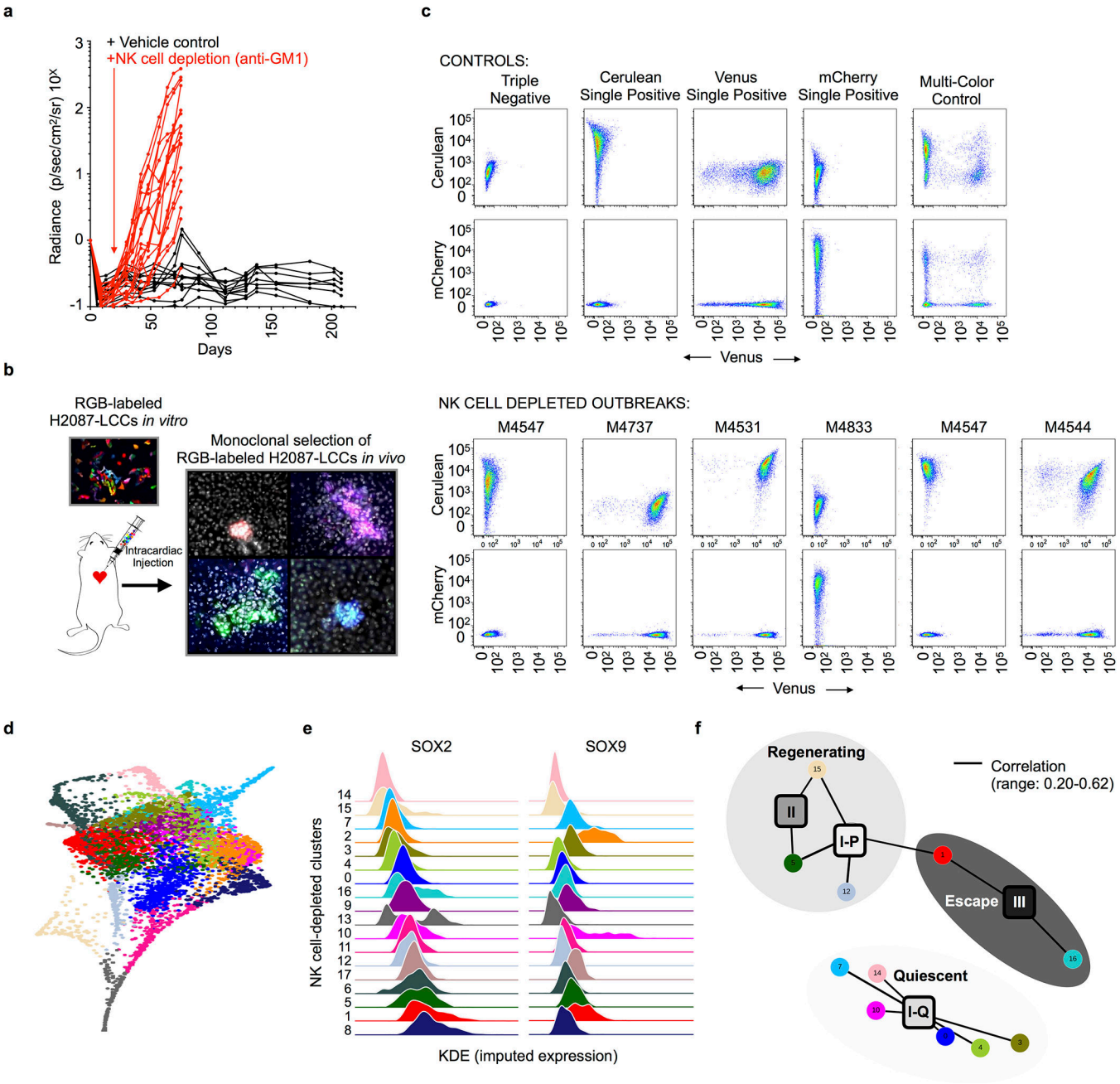
(bottom) denote patient metastatic clusters based on fraction of metastatic cells (>10%). (**b**) Tissue source, clusters (matching those depicted in **a**), and cell types (annotated as in Fig. 2), are visualized on a force directed layout (n = 3,786 cells). **c**, SOX2, SOX9 and DAPI immunofluorescence in three additional patient-matched primary tumor-metastasis pairs. Scale bars, 100 μm. **d**, Fraction of each Phenograph cluster detected per metastasis sample (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, fraction detected per n = 5 metastatic samples). **e,** Entropy of patient distribution in each cluster, computed with bootstrapping to correct for number of cells per cluster (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range, n = 100 random subsamples of data). Metastatic clusters are shaded by Phenograph cluster ID and clusters are ordered by average entropy. Metastatic clusters associated with alveolar epithelial progenitor signature (type III), and two clusters comprised predominantly of primary tumor cells treated with neo-adjuvant therapy are patient-specific. **f,** *Left,* Patient metastatic clusters ranked according to average lung epithelial development GO signature expression (n = 34 genes) less one gene. Each row shows metastatic cluster ranking for each left-out gene. *Right,* kernel density plot of imputed and normalized expression of each left-out gene.



Extended Data Fig. 9: Developmental stage-specific differential immune sensitivity extended.

a, Boxplots showing the average expression of NK cell-specific genes in TCGA lung adenocarcinoma patients stratified by SOX2 or SOX9 expression. NK cells are more abundant in SOX9^{high} tumors and conversely, less abundant in SOX2^{high} tumors (two-sided $p < 0.001$ based on Mann Whitney U test). Center line represents median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers. **b,** Cropped Western blots showing SOX2 and SOX9 protein levels upon DOX induction in the H2087-LCC model; no independent repeats were performed. Un-cropped Western blots are provided in Source Data. **c,** mRNA expression of Sox2 and Sox9 in bulk KP mouse LUAD derivatives[39], from primary tumors (circles) or metastases (x); red indicates the derivative is metatstatic. **d,** Endogenous expression of nuclear SOX9 enumerated by quantitative immunofluorescence in KP482T1 metastatic cells before and after co-culture with IL2-activated mouse NK cells. Fraction of SOX9 positive cells before and after NK cell co-culture are reported. Average of n = 3 technical replicates for each of 3 biological replicates. Between 5 and 9 locations were imaged and quantified per biological and technical replicate; resulting in quantitation of 4,534 cells before NK cell co-culture and 2,556 individual cells after NK cell co-culture. P-

values from paired two-tailed t-tests (n = 3 biological replicates, degrees of freedom = 2, t = 8.33, p = 0.01 for SOX9 single positive comparison). **e,** Relative mRNA expression of NK activating ligands in H2087-LCC cells with and without induction of SOX2 or SOX9 (n = 3 technical replicates; center values, mean; error bars, 95% confidence interval).



Extended Data Fig. 10: Clonality and phenotypic landscape of NK cell-depleted macrometastases.

a, Integrated radiance of H2087-LCC cells intracardially injected in mice, +/− anti-GM1 antibody treatment to deplete NK cells measured over time. **b,** Schematic illustrating trichromatic marking system implemented to assay the clonality of NK cell-depleted metastases. Fluorescence of trichromic reporter visualized in metastatic outbreaks generated in NSG mice lacking NK cells. **c,** FACS plots shows distribution of cells expressing Cerulean, Venus and mCherry per NK cell-depleted metastasis (lower) as compared to single and multi-color controls (upper); repeated independently for n = 6 NK cell-depleted

macrometastases. **d,** Force-directed layout of all single cells (n = 6,073 cells) transcriptionally profiled from 8 NK cell-depleted macrometastases colored by Phenograph cluster. **e**, Kernel density plot of the imputed expression of SOX2 and SOX9 in each NK cell-depleted Phenograph cluster; clusters are ranked by median of the SOX2 distribution. **f,** A bipartite graph representing genome-wide correlations across all common, variably expressed genes (n = 2,895; Methods) between each NK cell-depleted Phenograph cluster (n = 18,circular nodes) and each developmental state observed in human tumors (n = 4, square nodes, annotated in Fig. 3c). The Pearson correlation is computed across all categorical assignments between the two independent sets and edges link NK cell-depleted Phenograph clusters to human developmental for Pearson $R > 0.20$ and two-sided $p < 0.05$; edge width is scaled by the magnitude of the correlation (observed range: 0.20–0.62). Pearson correlation coefficients are also reported in Supplementary Table 2. Shading is used to highlight nodes assigned the three metastatic states detailed in Fig. 4.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the Thoracic Oncology Service, the Flow Cytometry Core, the Molecular Cytology Core, the Genomic Editing & Screening Core, and the Integrated Genomics Operation at MSKCC for their assistance. We especially thank the tissue donors at MSKCC for participating in this study. We thank J. Nainys and V. Kiseliovas for providing hands-on training of single cell library preparations. We thank E. Azizi and T. Nawy for their critical feedback on this manuscript and F. Bakhoum and V. Bakhoum for their support. This work was supported by NIH grants U54-CA209975 (JM and DP), P01-CA129243 (JM), DP1- HD084071 (DP), R01CA164729 (DP), CCSG P30 CA008748 Thompson Grant (JM), P30-CA008748 (MSKCC Molecular Cytology Core); DoD Innovator Award W81XWH-12-0074 (JM); and awards from the Burroughs Wellcome Fund Career Award at the Scientific Interface (AML), the Lung Cancer Research Foundation (AML), The Alan and Sandra Gerry Metastasis and Tumor Ecosystems Center at MSKCC (AML, JH), the Kirschstein-NRSA F30-CA220954 Predoctoral Fellowship (NRC) and by a T32-GM007729 Medical Scientist Training Program Grant (NRC).

REFERENCES

1. Beumer J & Clevers H Regulation and plasticity of intestinal stem cells during homeostasis and regeneration. Development 143, 3639–3649, doi:10.1242/dev.133132 (2016). [PubMed: 27802133]
2. Kumar PA et al. Distal airway stem cells yield alveoli in vitro and during lung regeneration following H1N1 influenza infection. Cell 147, 525–538, doi:10.1016/j.cell.2011.10.001 (2011). [PubMed: 22036562]
3. Vaughan AE et al. Lineage-negative progenitors mobilize to regenerate lung epithelium after major injury. Nature 517, 621–625, doi:10.1038/nature14112 (2015). [PubMed: 25533958]
4. Zuo W et al. p63(+)Krt5(+) distal airway stem cells are essential for lung regeneration. Nature 517, 616–620, doi:10.1038/nature13903 (2015). [PubMed: 25383540]
5. Zacharias WJ et al. Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor. Nature 555, 251–255, doi:10.1038/nature25786 (2018). [PubMed: 29489752]
6. Zaret KS & Grompe M Generation and regeneration of cells of the liver and pancreas. Science 322, 1490–1494, doi:10.1126/science.1161431 (2008). [PubMed: 19056973]
7. Kotton DN & Morrisey EE Lung regeneration: mechanisms, applications and emerging stem cell populations. Nat Med 20, 822–832, doi:10.1038/nm.3642 (2014). [PubMed: 25100528]
8. Murry CE & Keller G Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell 132, 661–680, doi:10.1016/j.cell.2008.02.008 (2008). [PubMed: 18295582]

9. Tata PR et al. Developmental History Provides a Roadmap for the Emergence of Tumor Plasticity. Dev Cell 44, 679–693 e675, doi:10.1016/j.devcel.2018.02.024 (2018). [PubMed: 29587142]
10. Massague J & Obenauf AC Metastatic colonization by circulating tumour cells. Nature 529, 298–306, doi:10.1038/nature17038 (2016). [PubMed: 26791720]
11. Klein AM et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201, doi:10.1016/j.cell.2015.04.044 (2015). [PubMed: 26000487]
12. Zilionis R et al. Single-cell barcoding and sequencing using droplet microfluidics. Nat Protoc 12, 44–73, doi:10.1038/nprot.2016.154 (2017). [PubMed: 27929523]
13. Azizi E et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell, doi:10.1016/j.cell.2018.05.060 (2018).
14. van Dijk D et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell, doi:10.1016/j.cell.2018.05.061 (2018).
15. Levine JH et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 162, 184–197, doi:10.1016/j.cell.2015.05.047 (2015). [PubMed: 26095251]
16. Malladi S et al. Metastatic Latency and Immune Evasion through Autocrine Inhibition of WNT. Cell 165, 45–60, doi:10.1016/j.cell.2016.02.025 (2016). [PubMed: 27015306]
17. Tirosh I et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 539, 309–313, doi:10.1038/nature20123 (2016). [PubMed: 27806376]
18. Lavin Y et al. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. Cell 169, 750–765, doi:10.1016/j.cell.2017.04.014 (2017). [PubMed: 28475900]
19. Shen R & Seshan VE FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res 44, e131, doi:10.1093/nar/gkw520 (2016). [PubMed: 27270079]
20. Cheng DT et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn 17, 251–264, doi:10.1016/j.jmoldx.2014.12.006 (2015). [PubMed: 25801821]
21. Bremnes RM et al. The Role of Tumor-Infiltrating Lymphocytes in Development, Progression, and Prognosis of Non-Small Cell Lung Cancer. J Thorac Oncol 11, 789–800, doi:10.1016/j.jtho.2016.01.015 (2016). [PubMed: 26845192]
22. Morrisey EE & Hogan BL Preparing for the first breath: genetic and cellular mechanisms in lung development. Dev Cell 18, 8–23, doi:10.1016/j.devcel.2009.12.010 (2010). [PubMed: 20152174]
23. Du YN et al. Lung Gene Expression Analysis (LGEA): an integrative web portal for comprehensive gene expression data analysis in lung development. Thorax 72, 481–484, doi:10.1136/thoraxjnl-2016-209598 (2017). [PubMed: 28070014]
24. Du YN, Guo MZ, Whitsett JA & Xu Y 'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung. Thorax 70, 1092–1094, doi:10.1136/thoraxjnl-2015-207035 (2015). [PubMed: 26130332]
25. Treutlein B et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375, doi:10.1038/nature13173 (2014). [PubMed: 24739965]
26. Guha A et al. Neuroepithelial body microenvironment is a niche for a distinct subset of Clara-like precursors in the developing airways. Proc Natl Acad Sci U S A 109, 12592–12597, doi:10.1073/pnas.1204710109 (2012). [PubMed: 22797898]
27. Haghverdi L, Buttner M, Wolf FA, Buettner F & Theis FJ Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods 13, 845–848, doi:10.1038/nmeth.3971 (2016). [PubMed: 27571553]
28. Setty M et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol 34, 637–645, doi:10.1038/nbt.3569 (2016). [PubMed: 27136076]
29. Jobe AH, Whitsett J & Abman SH Fetal & neonatal lung development : clinical correlates and technologies for the future. (Cambridge University Press, 2015).
30. Nakamura N et al. Identification of tumor markers and differentiation markers for molecular diagnosis of lung adenocarcinoma. Oncogene 25, 4245–4255, doi:10.1038/sj.onc.1209442 (2006). [PubMed: 16491115]

31. Smith BA et al. A Human Adult Stem Cell Signature Marks Aggressive Variants across Epithelial Cancers. Cell Rep 24, 3353–3366 e3355, doi:10.1016/j.celrep.2018.08.062 (2018). [PubMed: 30232014]
32. Fetal and Neonatal Lung Development Clinical Correlates and Technologies for the Future. (Cambridge University Press, 2016).
33. Niakan KK et al. Sox17 promotes differentiation in mouse embryonic stem cells by directly regulating extraembryonic gene expression and indirectly antagonizing self-renewal. Genes Dev 24, 312–326, doi:10.1101/gad.1833510 (2010). [PubMed: 20123909]
34. Seguin CA, Draper JS, Nagy A & Rossant J Establishment of endoderm progenitors by SOX transcription factor expression in human embryonic stem cells. Cell Stem Cell 3, 182–195, doi:10.1016/j.stem.2008.06.018 (2008). [PubMed: 18682240]
35. Okubo T, Knoepfler PS, Eisenman RN & Hogan BL Nmyc plays an essential role during lung development as a dosage-sensitive regulator of progenitor cell proliferation and differentiation. Development 132, 1363–1374, doi:10.1242/dev.01678 (2005). [PubMed: 15716345]
36. Rawlins EL, Clark CP, Xue Y & Hogan BL The Id2+ distal tip lung epithelium contains individual multipotent embryonic progenitor cells. Development 136, 3741–3745, doi:10.1242/dev.037317 (2009). [PubMed: 19855016]
37. Gyorffy B, Surowiak P, Budczies J & Lanczky A Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. PLoS One 8, e82241, doi:10.1371/journal.pone.0082241 (2013). [PubMed: 24367507]
38. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550, doi:10.1038/nature13385 (2014). [PubMed: 25079552]
39. Winslow MM et al. Suppression of lung adenocarcinoma progression by Nkx2–1. Nature 473, 101–104, doi:10.1038/nature09881 (2011). [PubMed: 21471965]
40. Jung H, Hsiung B, Pestal K, Procyk E & Raulet DH RAE-1 ligands for the NKG2D receptor are regulated by E2F transcription factors, which control cell cycle entry. J Exp Med 209, 2409–2422, doi:10.1084/jem.20120565 (2012). [PubMed: 23166357]
41. Long EO Negative signaling by inhibitory receptors: the NK cell paradigm. Immunol Rev 224, 70–84, doi:10.1111/j.1600-065X.2008.00660.x (2008). [PubMed: 18759921]
42. Er EE et al. Pericyte-like spreading by disseminated cancer cells activates YAP and MRTF for metastatic colonization. Nat Cell Biol 20, 966–978, doi:10.1038/s41556-018-0138-8 (2018). [PubMed: 30038252]
43. Weber K et al. RGB marking facilitates multicolor clonal cell tracking. Nat Med 17, 504–509, doi:10.1038/nm.2338 (2011). [PubMed: 21441917]
44. Finak G et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 16, doi:ARTN 278 10.1186/s13059–015-0844–5 (2015).
45. Nowotschin S et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. Nature 569, 361–+, doi:10.1038/s41586-019-1127-1 (2019). [PubMed: 30959515]
46. Mathelier A et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res 42, D142–147, doi:10.1093/nar/gkt997 (2014). [PubMed: 24194598]
47. Mathelier A et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 44, D110–115, doi:10.1093/nar/gkv1176 (2016). [PubMed: 26531826]
48. Halko N MP, Tropp JA. Finding structure with randomness: Probabilistic algorithms for constructing apprxoimate matrix decompositions. ARXIV (2009).
49. Valle S, Li WH & Qin SJ Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods. Ind Eng Chem Res 38, 4389–4401, doi:DOI 10.1021/ie990110i (1999).
50. van der Maaten L & Hinton G Visualizing Data using t-SNE. J Mach Learn Res 9, 2579–2605 (2008).

51. Jacomy M, Venturini T, Heymann S & Bastian M ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 9, e98679, doi:10.1371/journal.pone.0098679 (2014). [PubMed: 24914678]

52. Liberzon A et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740, doi:10.1093/bioinformatics/btr260 (2011). [PubMed: 21546393]

53. Rock JR et al. Basal cells as stem cells of the mouse trachea and human airway epithelium. Proc Natl Acad Sci U S A 106, 12771–12775, doi:10.1073/pnas.0906850106 (2009). [PubMed: 19625615]

54. Li S et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nature Immunology 15, 195–204, doi:10.1038/ni.2789 (2014). [PubMed: 24336226]

55. Abbas AR et al. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun 6, 319–331, doi:10.1038/sj.gene.6364173 (2005). [PubMed: 15789058]

56. Novershtern N et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144, 296–309, doi:10.1016/j.cell.2011.01.004 (2011). [PubMed: 21241896]

57. Jeffrey KL et al. Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nat Immunol 7, 274–283, doi:10.1038/ni1310 (2006). [PubMed: 16474395]

58. Mor-Vaknin N, Punturieri A, Sitwala K & Markovitz DM Vimentin is secreted by activated macrophages. Nat Cell Biol 5, 59–63, doi:10.1038/ncb898 (2003). [PubMed: 12483219]

59. Zepp JA et al. Distinct Mesenchymal Lineages and Niches Promote Epithelial Self-Renewal and Myofibrogenesis in the Lung. Cell 170, 1134–1148 e1110, doi:10.1016/j.cell.2017.07.034 (2017). [PubMed: 28886382]

METHODS-ONLY REFERENCES

60. Lee JH et al. Anatomically and Functionally Distinct Lung Mesenchymal Populations Marked by Lgr5 and Lgr6. Cell 170, 1149–1163, doi:10.1016/j.cell.2017.07.028 (2017). [PubMed: 28886383]

61. Xia H et al. Calcium-binding protein S100A4 confers mesenchymal progenitor cell fibrogenicity in idiopathic pulmonary fibrosis. J Clin Invest 127, 2586–2597, doi:10.1172/JCI90832 (2017). [PubMed: 28530639]

62. Degryse AL et al. Repetitive intratracheal bleomycin models several features of idiopathic pulmonary fibrosis. Am J Physiol Lung Cell Mol Physiol 299, L442–452, doi:10.1152/ajplung.00026.2010 (2010). [PubMed: 20562227]

63. Tanjore H et al. Contribution of epithelial-derived fibroblasts to bleomycin-induced lung fibrosis. Am J Respir Crit Care Med 180, 657–665, doi:10.1164/rccm.200903-0322OC (2009). [PubMed: 19556518]

64. Lawson WE et al. Characterization of fibroblast-specific protein 1 in pulmonary fibrosis. Am J Respir Crit Care Med 171, 899–907, doi:10.1164/rccm.200311-1535OC (2005). [PubMed: 15618458]

65. Li ZH, Dulyaninova NG, House RP, Almo SC & Bresnick AR S100A4 regulates macrophage chemotaxis. Mol Biol Cell 21, 2598–2610, doi:10.1091/mbc.E09-07-0609 (2010). [PubMed: 20519440]

66. Moore KW, de Waal Malefyt R, Coffman RL & O'Garra A Interleukin-10 and the interleukin-10 receptor. Annu Rev Immunol 19, 683–765, doi:10.1146/annurev.immunol.19.1.683 (2001). [PubMed: 11244051]

67. Priceman SJ et al. Targeting distinct tumor-infiltrating myeloid cells by inhibiting CSF-1 receptor: combating tumor evasion of antiangiogenic therapy. Blood 115, 1461–1471, doi:10.1182/blood-2009-08-237412 (2010). [PubMed: 20008303]

68. Lambrechts D et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 24, 1277–1289, doi:10.1038/s41591-018-0096-5 (2018). [PubMed: 29988129]

69. Zilionis R et al. Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. Immunity 50, 1317–1334 e1310, doi:10.1016/j.immuni.2019.03.009 (2019). [PubMed: 30979687]

70. Coifman RR et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci U S A 102, 7426–7431, doi:10.1073/pnas.0500334102 (2005). [PubMed: 15899970]

71. Setty M et al. Characterization of cell fate probabilities in single-cell data with Palantir. Nat Biotechnol 37, 451–460, doi:10.1038/s41587-019-0068-4 (2019). [PubMed: 30899105]

72. Davoli T, Uno H, Wooten EC & Elledge SJ Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science 355, doi:10.1126/science.aaf8399 (2017).

73. Yarilin D et al. Machine-based method for multiplex in situ molecular characterization of tissues by immunofluorescence detection. Sci Rep 5, 9534, doi:10.1038/srep09534 (2015). [PubMed: 25826597]

74. Otsu N Threshold Selection Method from Gray-Level Histograms. Ieee T Syst Man Cyb 9, 62–66, doi:Doi 10.1109/Tsmc.1979.4310076 (1979).

75. He K, Gkioxari G, Dollar P & Girshick R Mask R-CNN. IEEE Trans Pattern Anal Mach Intell, doi:10.1109/TPAMI.2018.2844175 (2018).

76. Loken MR, Parks DR & Herzenberg LA Two-color immunofluorescence using a fluorescence-activated cell sorter. J Histochem Cytochem 25, 899–907, doi:10.1177/25.7.330738 (1977). [PubMed: 330738]

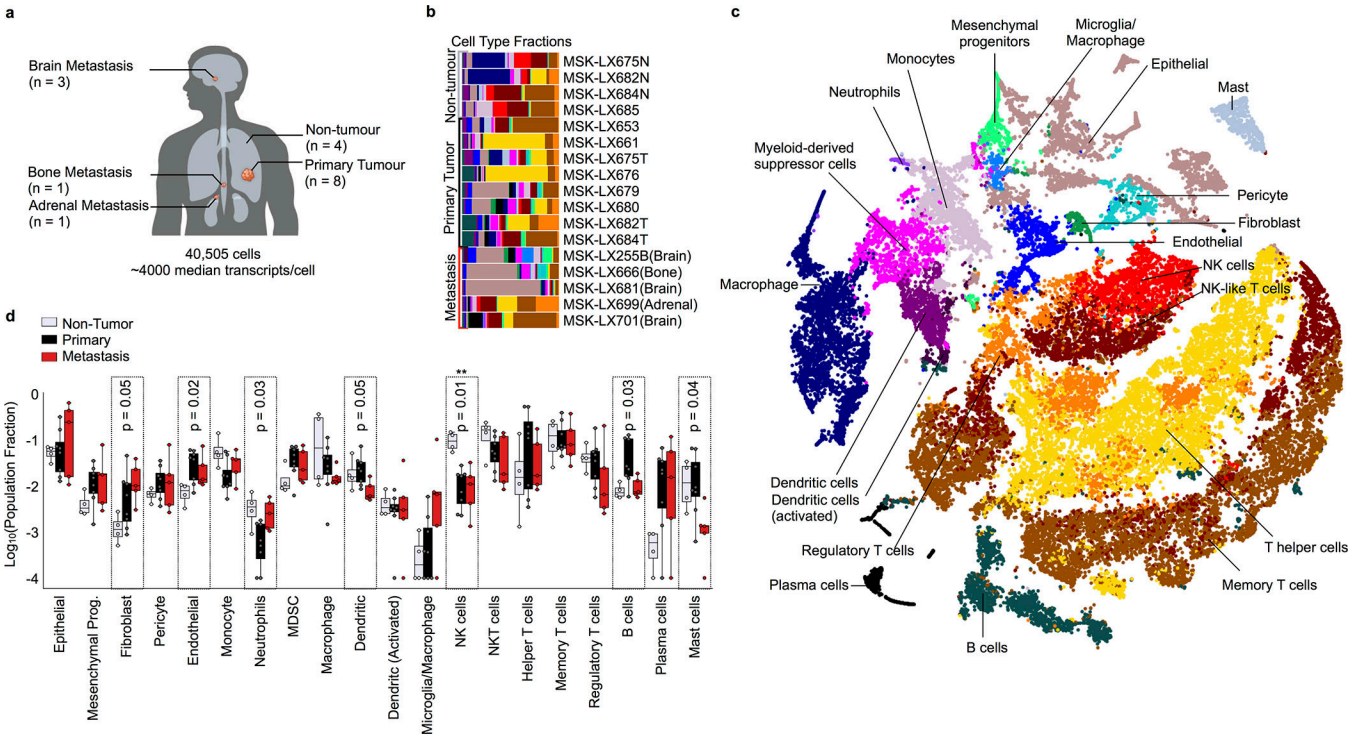


Figure 1. The single-cell transcriptional landscape of human lung adenocarcinoma.

a, Patient tissues profiled (metadata summarized in Extended Data Fig. 1). **b**, Cell type fractions detected per sample, color coded as in **c**. **c**, t-SNE projection of the complete atlas of normal lung, primary tumour and metastatic LUAD colored by cell type; includes carcinoma and non-tumour epithelium, as well as immune and other stromal cell types within the tumours (n = 40,505 cells). **d**, Cell-type abundances differ between normal, primary and metastatic sites (n = 17 patient samples; center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers). Significant differences in cell type abundance are highlighted (Kruskal-Wallis rank test).

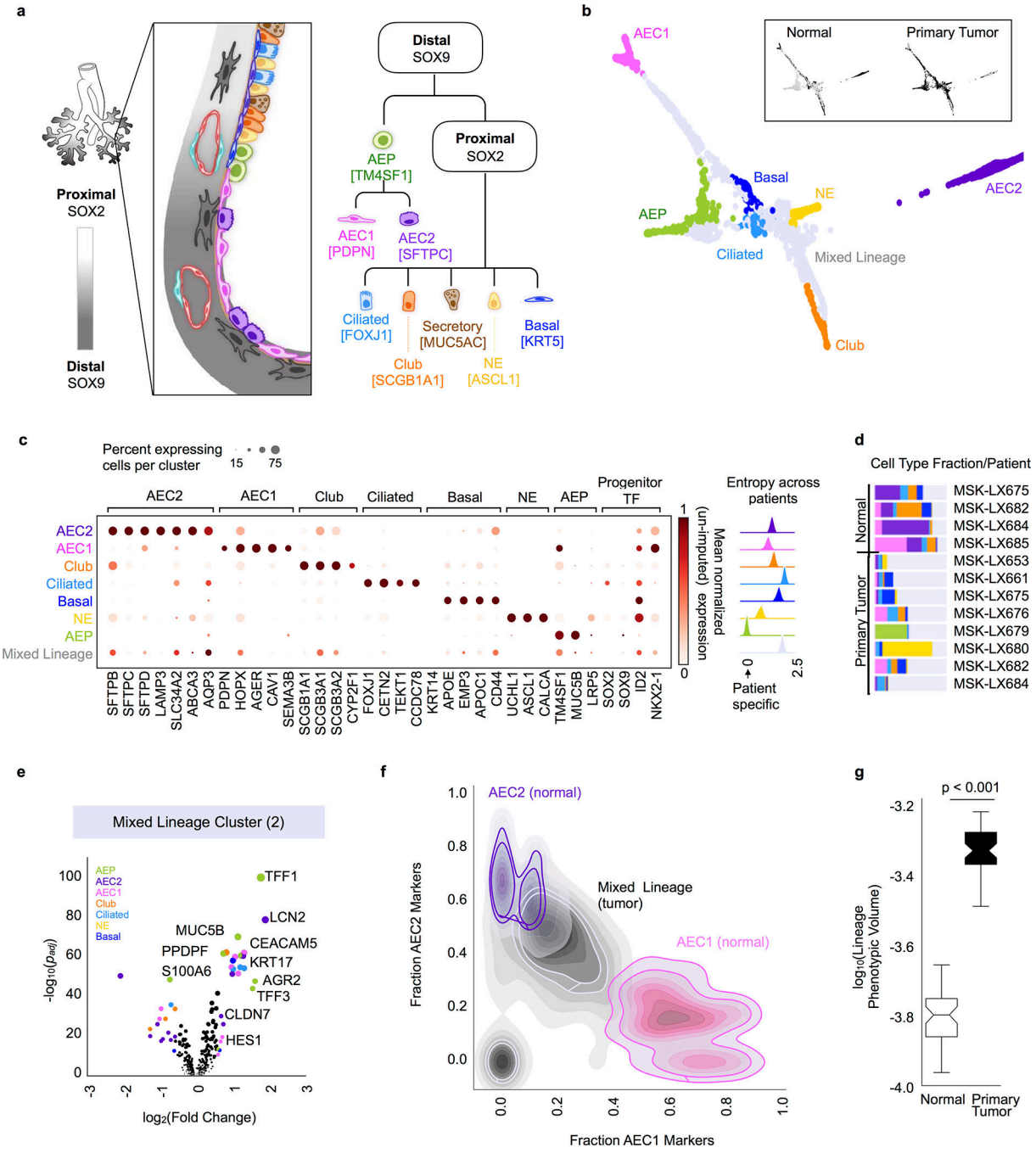


Figure 2. Regenerative and mixed lineages in primary tumours.

a, Key epithelial cell types and associated transcription factors (in brackets) implicated in lung development and regeneration[29]. Specialized cell types of the SOX9-specified distal lung include type I alveolar epithelial cells (AEC1) that perform gas exchange, and type II alveolar epithelial cells (AEC2) that secrete surfactants. The SOX2-specified proximal upper airway includes secretory, neuroendocrine (NE) and ciliated cell types that predominantly serve barrier functions. Infiltrating stromal cell types that support lung epithelial development and are detected by scRNA-seq are also illustrated. **b**, Force-directed layout of

normal and primary tumour epithelial cells colored by annotated lineage (n = 2,140 cells). Tumour cells that concomitantly express multiple cell type markers (mixed lineage) are in grey. *Inset,* same layout colored (in black) by source of epithelium. **c**, *Left,* Relative frequency of cells expressing each canonical lineage marker (any counts detected in un-imputed data) and the average un-imputed expression of each gene (z-normalized across epithelial cell types) for all expressing cells in a given cluster. *Right,* Kernel density plot depicting entropy of cell mixing across all patients for each cell type, computed with bootstrapping to correct for number of cells in each cluster (n = 100 random subsamples of data). High entropy indicates most similar cells come from a well-mixed set of patient samples; low entropy indicates most similar cells come from the same patient sample. **d**, Fraction of each cell type detected per patient sample (colors as in **b**). **e**, Top DEGs for a representative mixed lineage cluster (Cluster 2, n = 183 cells) compared to all other cells computed using MAST[44], indicating enrichment of AEC1, AEC2, club, ciliated, basal and AEP markers (volcano plots supporting other mixed-lineage clusters are in Extended Data Fig. 7a). Lineage-specific DEGs are colored by associated cell type (as in **b**), diameter proportional to $-log_{10}(p_{adj})$ for genes with *fold change* > 1.5 and $p_{adj} < 0.05$ (see Supplementary Table 1 for lineage-specific genes). **f,** 2D cell density plot showing fraction of AEC1 and AEC2 lineage markers per cell in normal lung alveolar cells compared to tumour cells from Clusters 0 and 1. Only markers with normalized un-imputed expression in the top quartile (per gene, across all cells) are plotted. The overlapping distributions are shaded by cell type. **g,** Lineage phenotypic volume (Methods) of epithelium derived from primary tumours compared to normal lung, showing significant expansion of lineage gene-gene covariate structure in primary tumours (two-sided Mann-Whitney rank test, $p < 0.001$). n = 50 random subsamples of the data each (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers).

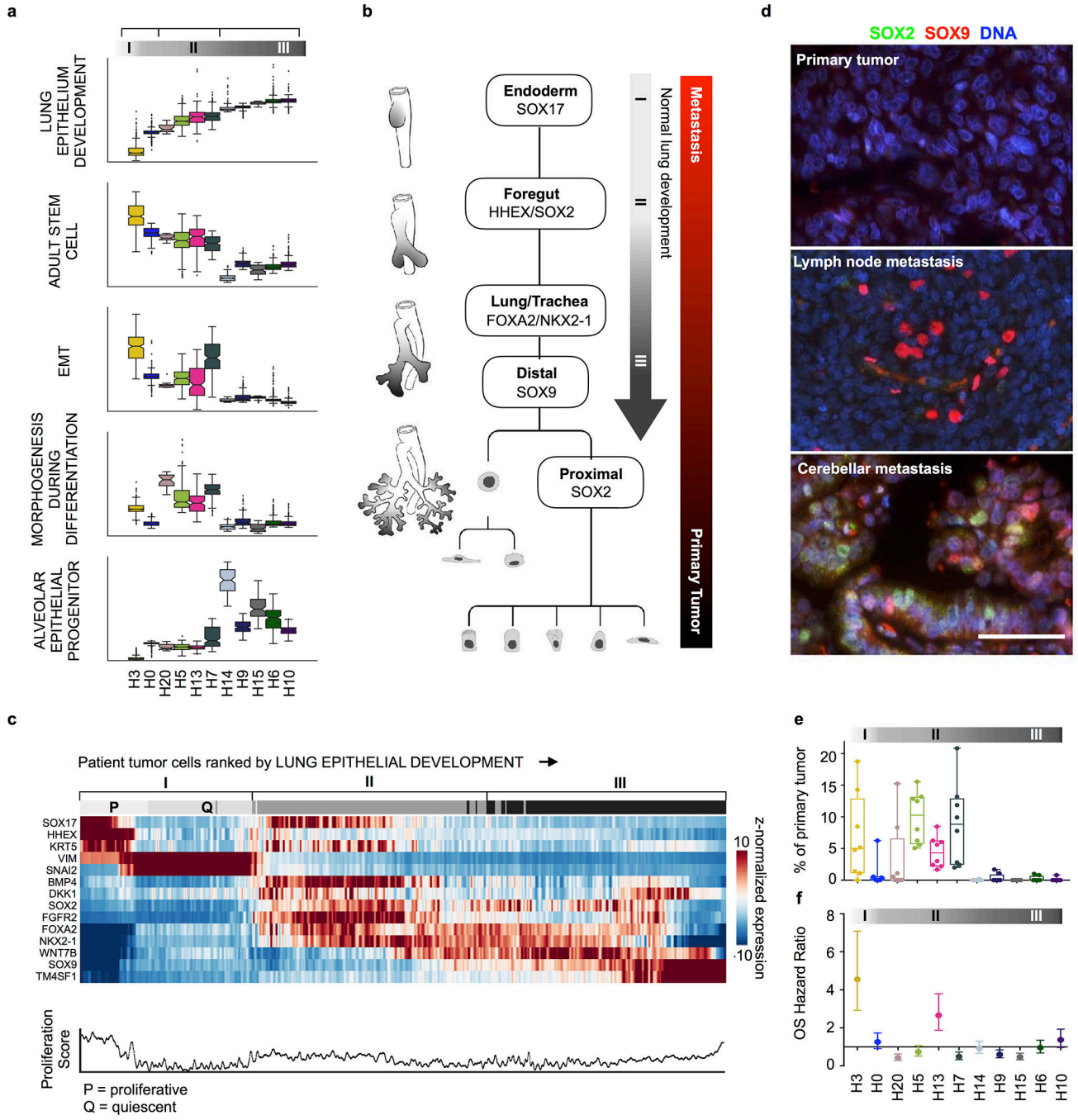


Figure 3. Metastases exhibit a continuum of stem to lung epithelial progenitor states.

a, For each patient metastatic cluster, boxplots indicate the cellular distribution of the average imputed and normalized expression of five key gene signatures associated with lung development (see Supplementary Table 2 for the signatures). Clusters are ranked by average expression of the lung epithelial signature; cluster color and labels are as in Extended Data Fig. 8a–b. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers. Cells from type I clusters (n = 813) are associated with increased expression of an adult stem cell signature (two-sided Mann-Whitney U-test: U =

281595, p = 7e-81). Cells from type I and II clusters (n = 1,408) show increased expression of genes involved in epithelial-mesenchymal transition (EMT) (U = 215794, p = 1e-115). Cells from type II intermediate clusters (n = 595) were specifically enriched for regenerative pathways related to morphogenesis and specification of the respiratory endoderm (U = 332492, p =2e-25). Whereas cells from type III clusters (n = 756) exhibited the highest level of alveolar epithelial progenitor programs (U = 40174, p = 2e-276) and decreased expression of adult stem cell genes (U = 229739, p = 6e-106). **b**, Relationship between type I-III assignments and key transcription factors specifying stem and lung epithelial progenitors in the canonical model of lung morphogenesis[29]. **c**, Imputed and normalized expression of key transcription factors specifying stem and lung epithelial progenitors (rows) for all individual tumour cells, ranked by average expression of the lung epithelial development GO signature (Supplementary Table 2) in ascending order from left to right. Expression of each transcription factor was z-normalized across all cells and smoothed using a 20-cell moving average widow. Clustering applied directly to this matrix (Methods) assigned each cell to a proliferating (P) or quiescent (Q) stem-like state (type I), regenerative state (type II), or a *SOX9*^{high} alveolar epithelial progenitor state (type III) (top row). *Bottom*, average expression of three canonical proliferation markers across ranked tumour cells (Methods). **d**, SOX2 and SOX9 immunofluorescence and DAPI in matched primary tumour, lymph node and cerebellar metastases from representative patient (repeated in n = 4 independent matched patients with similar results). Scale bar, 50 μm. Additional patients in Extended Data Figure 8c. **e,** Percentage of each metastatic cluster detected per primary tumour. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; each point is a primary tumour (n = 8 independent patient samples). **f,** Hazard ratio (HR) with 95% confidence intervals for overall survival (OS), computed between lowest and highest quartiles for n = 673 LUAD patients[37] (Methods). HR > 1 is poorly prognostic; HR < 1 indicates improved OS and any CI crossing the line at 1 is not significant.

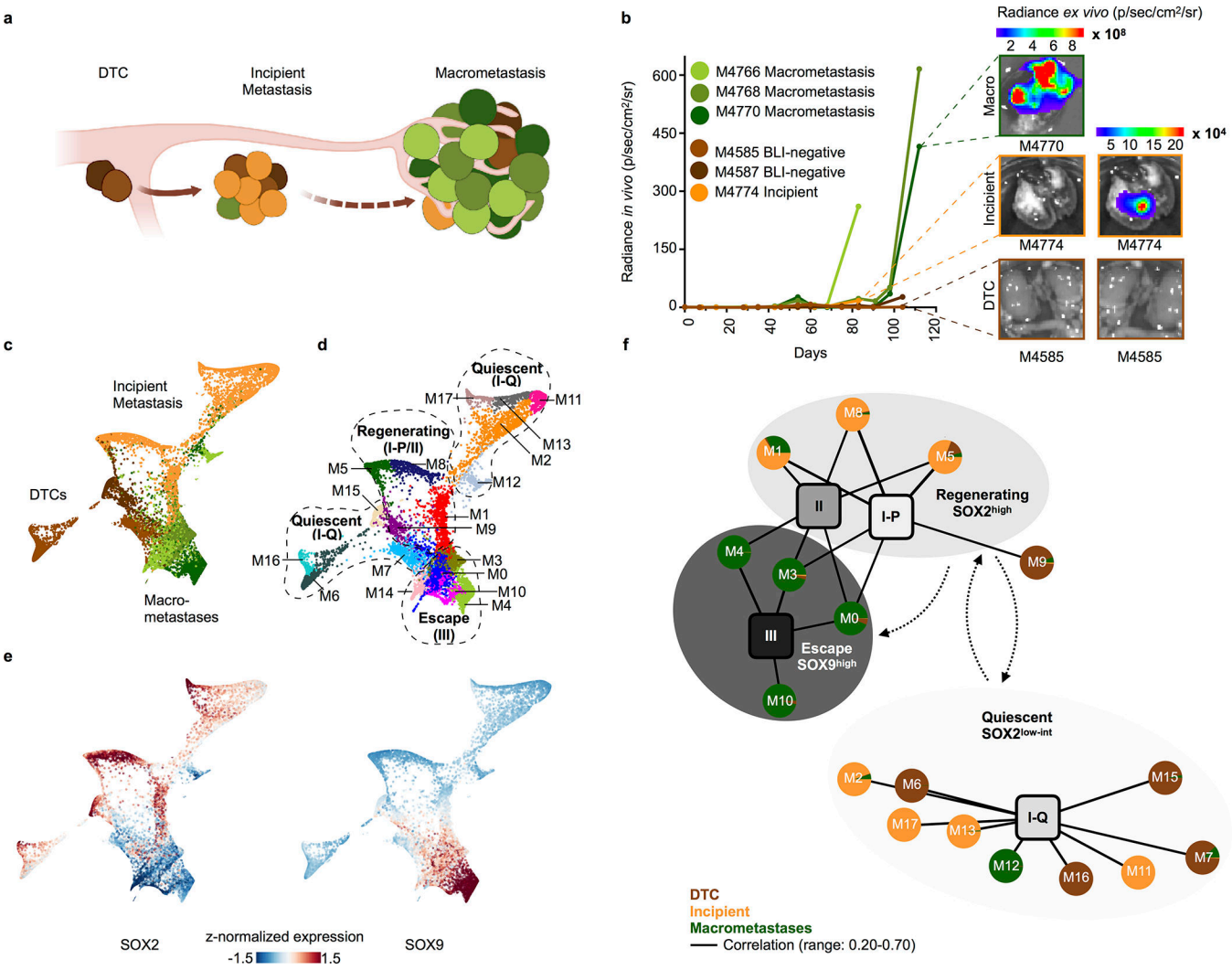


Figure 4. Developmental continuum in a mouse model of metastatic escape.

a, Xenograft model of metastatic escape, illustrating the three stages from which single cells were sampled. **b**, BLI growth curves (measured *ex vivo* for up to 120 days) for all sampled xenograft metastases, including *ex vivo* BLI images acquired at tissue harvest. **c-d,** Force-directed layout of all metastatic tumour cells (n = 8,748 cells) isolated from 6 mice colored by (**c**) source and **(d)** Phenograph cluster. Clusters were grossly assigned to one of three metastatic states: *Quiescen*t, correlated with a non-proliferating stem-like state (type I-Q); *Regenerating*, correlated with proliferating stem (type I-P) and the regenerative (type II) state; and *Escape*, which is highly concordant with *SOX9*^{high} alveolar epithelial progenitors (type III) (see Methods). **e,** Force-directed layout (as in **c,d**) of all xenograft tumour cells colored by imputed, z-normalized *SOX2* and *SOX9* expression. **f,** Bipartite graph representing genome-wide correlations across all common, variably expressed genes (n = 2,096, Methods) between the 18 mouse clusters (circular nodes) and 4 developmental human tumour states (square nodes, annotated in Fig. 3c). Pie charts within circular nodes represent mouse cell sources. Edges link mouse clusters and human states with genome-wide Pearson $R > 0.20$ and two-sided $p < 0.05$; edge width is proportional to the correlation magnitude

(see Supplementary Table 2 for exact values). Dotted arrows suggest temporal ordering between metastatic states, based on the three stages from which cells were isolated and profiled (according to BLI signature).

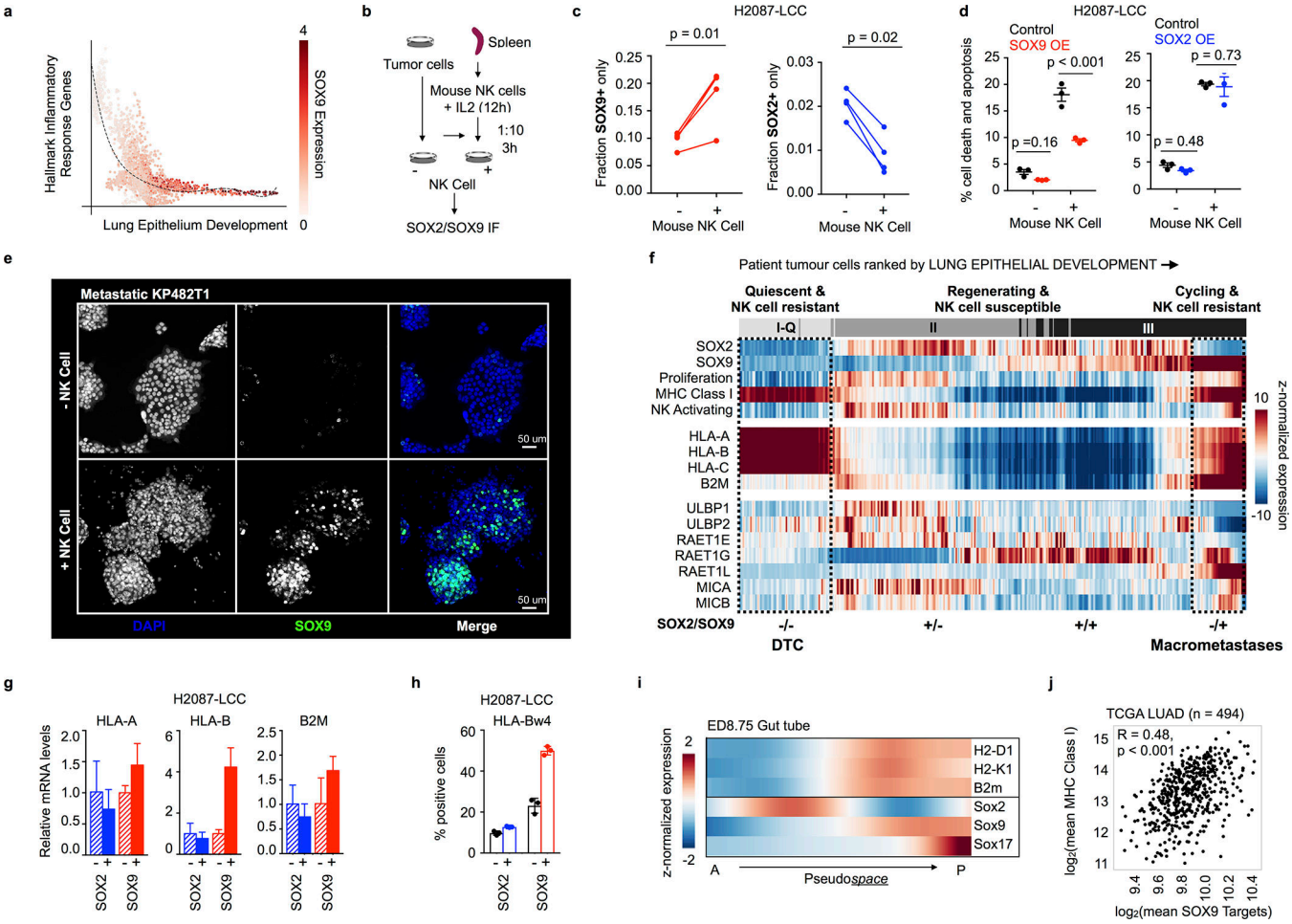


Figure 5. Developmental stage-specific differential immune sensitivity.

a, Average imputed and normalized Hallmark Inflammatory Response gene expression (Supplementary Table 2) across all patient-derived tumour cells, ranked by average lung epithelial development score. Each dot represents a cell colored by its imputed and normalized *SOX9* transcript counts. **b,** NK cell co-culture assay. Tumour cells were cultured alone or with IL2-activated mouse NK cells at a 1:10 ratio for 3 h, and endogenous SOX2 and SOX9 were detected by immunofluorescence (IF). **c,** Fraction of H2087-LCC cells exclusively positive for either nuclear SOX9 or SOX2 before and after co-culture (average of 3 technical replicates for each of 4 independent experiments; paired two-sided t-tests, 3 degrees of freedom). A total of 134,946 cells before co-culture and 32,602 cells after co-culture were quantified. **d,** Percentage of cell death and apoptosis in H2087-LCC cells measured by flow cytometry before and after co-culture in the context of inducible SOX2 and SOX9 over-expression (center line, mean; whiskers, SEM; points, 3 independent experiments; unpaired two-sided t-test, 8 degrees of freedom). **e,** SOX9 IF in KP482T1 mouse metastatic tumour cells before and after 3-h co-culture of tumour and IL2-activated mouse NK cells at a ratio of 1:10 (repeated in 3 independent experiments with similar results). **f,** Expression of transcription factors specifying stem and lung epithelial progenitors, MHC Class I markers of self, and NK activating ligands across patient-derived tumour cells assigned to type I-Q, II or III developmental stages (top row, as in Fig. 3), in

patient primary tumours and metastases. Proliferation refers to mean *PCNA*, *MKI67* and *MCM2* expression per cell. MHC Class I and NK activating show the average expression of their associated genes, visualized individually below. For each gene, imputed expression was z-normalized across all cells and smoothed using a 20-cell moving average widow. Dashed boxes indicate association with spontaneous micro- or macro-metastases observed in our xenograft model. Bottom, SOX2/SOX9 status. **g,** Relative expression of MHC Class I genes important for NK cell evasion in H2087-LCC cells with and without SOX2 or SOX9 induction, measured by RT-PCR (n = 3 technical replicates; center values, mean; error bars, 95% confidence interval). **h,** Cells positive for HLA Class I Bw4 surface protein, measured by flow cytometry (n = 3 independent experiments; center values, mean; error bars, 95% confidence interval; points, all measured data). **i,** Imputed expression of MHC Class I markers of self and Sox transcription factors specifying stem and lung epithelial progenitors in the D8.75 mouse gut tube, showing spatial segregation of *Sox2* and *Sox9* lineages in cells ranked by their pseudospace ordering[45]. A, anterior; P, posterior. **j,** Correlation between average *SOX9* target gene expression, predicted using motifs from the JASPAR Predicted Transcription Factor targets dataset (Supplementary Table 2)[46,47] and the average expression of all MHC Class I genes across n = 510 TCGA LUAD patients (Pearson R = 0.48 and two-sided $p < 0.001$ to test for non-correlation). Outliers defined as 1.5X the interquartile range less than Q1 or greater than Q3 (n = 16) are removed from the scatter plot.

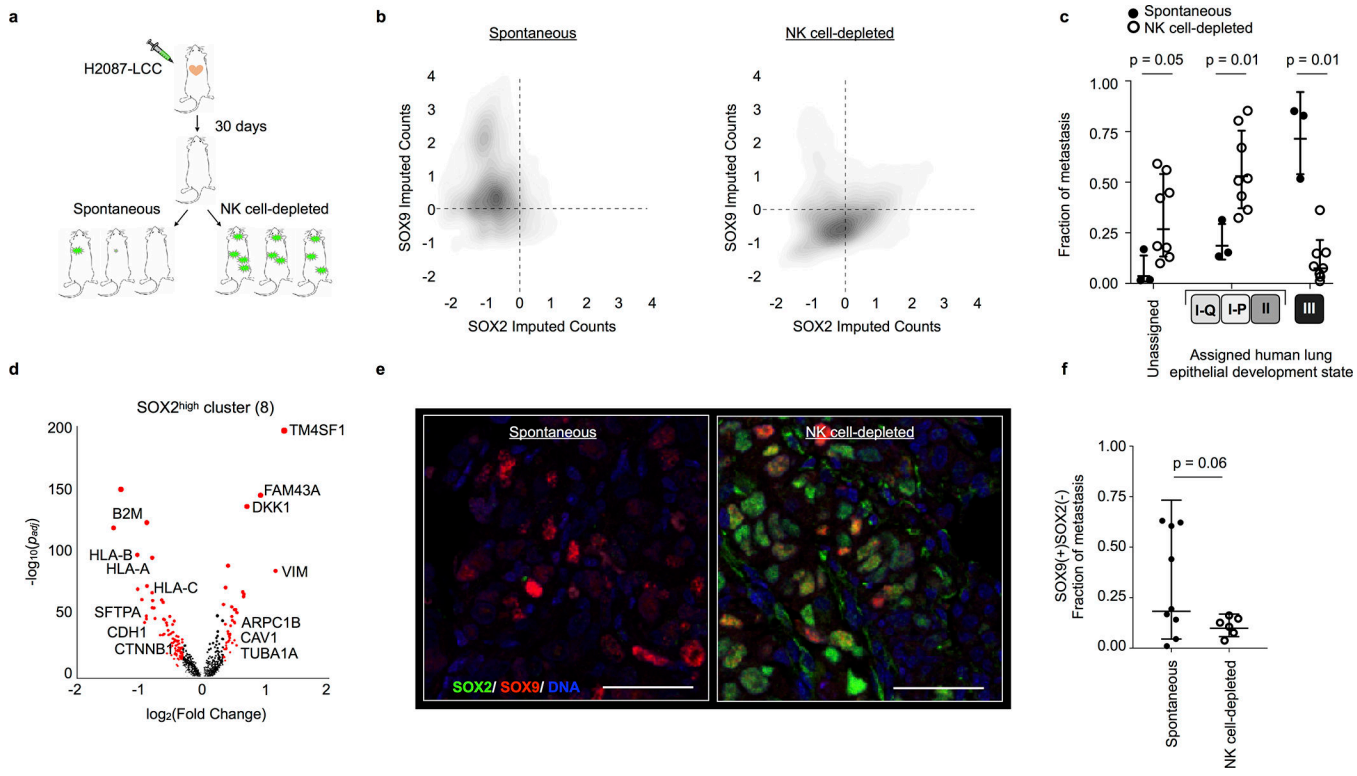


Figure 6. NK cell-dependent pruning limits the phenotypic expansion of metastasis-initiating cells.

a, *in vivo* NK cell perturbation assay in mice harboring latent metastasis-initiating cells. **b,** 2D cell density plot of z-normalized *SOX2* and *SOX9* imputed expression in H2087-LCC cells isolated from macrometastases +/− NK cell depletion, as determined by scRNA-seq. **c,** Fraction of type I/II and type III cells detected in spontaneous versus NK cell-depleted macrometastases (3 spontaneous macrometastases harvested from n = 3 independent mice and 8 NK cell-depleted macrometastases harvested from n = 5 independent mice; center line, geometric mean; whiskers, geometric s.d.; points, all measured data; two-sided Mann-Whitney rank test). Cell types are assigned by significant correlation with patient tumour states (Pearson $R > 0.20$ and two-sided $p < 0.05$ to test for non-correlation; as in Fig. 4f). **d,** Top DEG for NK cell-depleted cluster with highest *SOX2* expression (Phenograph cluster 8, n = 322 cells, see Extended Data Fig. 10e) compared to all other cells, computed using MAST[44]. DEGs are red, with diameter proportional to $-log_{10}(p_{adj})$ for genes with *fold change* > 1.5 and $p_{adj} < 0.05$. **e,** SOX2 and SOX9 immunofluorescence in a representative spontaneous and NK cell-depleted macrometastasis (n = 15 macrometastases evaluated, nuclear SOX9 expression summarized in **f**). Scale bars, 50 μm. **f,** Nuclear SOX2 and SOX9 single-positive, double-positive, and negative cell fractions quantified per macrometastatic lesion (n = 11,376 single cells quantified, fraction of metastases reported across n=15 lesions including lung, bone, kidney, and soft connective tissues harvested from 7 mice). 5 representative 20X frames were evaluated per lesion. SOX9 single-positive cells were enriched in spontaneous as compared to NK-cell-depleted macrometastases (n = 15 independent macrometastases, p = 0.06, one-sided Mann-Whitney rank test); abundance of

other cell types was not significantly altered (data not shown). Center line, geometric mean; whiskers, geometric s.d.; points, all measure data.