An atlas of epithelial cell states and plasticity in lung adenocarcinoma

Guangchun Han*[1], Ansam Sinjab*[2], Warapen Treekitkarnmongkol[2], Wei Lu[2], Alejandra G. Serrano[2], Sharia D. Hernandez[2], Zahraa Rahal[2], Xuanye Cao[1], Enyu Dai[1], Lorena I. Gomez-Bolanos[2], Edwin R. Parra[2], Marcelo Vailati Negrao[3], Tina Cascone[3], Boris Sepesi[4], Seyed Javad Moghaddam[5], Jichao Chen[5], Steven M. Dubinett[6], Paul Scheet[7], Junya Fujimoto[2], Luisa M. Solis[2], Ignacio I. Wistuba[2], Christopher S. Stevenson[8], Avrum E. Spira[8,9], Linghua Wang[#1,10], Humam Kadara[#2,10]

Departments of [1]Genomic Medicine, [2]Translational Molecular Pathology, [3]Thoracic, Head and Neck Medical Oncology, [4]Cardiovascular and Thoracic Surgery, and [5]Pulmonary Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX. [6]Department of Medicine, The University of California Los Angeles, Los Angeles, CA. [7]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX. [8]Lung Cancer Initiative at Johnson and Johnson, Boston, MA. [9]Section of Computational Biomedicine, School of Medicine, Boston University, Boston, MA. [10]The University of Texas Graduate School of Biomedical Sciences, Houston, TX.

* These authors contributed equally

[#] **Correspondence**: Humam Kadara, PhD, Email: hkadara@mdanderson.org ; Linghua Wang, MD, PhD: lwang22@mdanderson.org.

Abstract

Understanding cellular processes underlying early lung adenocarcinoma (LUAD) development is needed to devise intervention strategies[1]. Here, we studied 246,102 single epithelial cells from 16 early-stage LUADs and 47 matched normal lung samples. Epithelial cells comprised diverse normal alveolar and airway lineages as well as cancer cell populations. Diversity among cancer cells was strongly linked to LUAD patient-specific oncogenic drivers. *KRAS*-mutant cancer cells were unique in their transcriptional features, strikingly reduced differentiation, low levels of DNA copy number changes, and increased variability amongst the cells themselves. The local niche of LUADs, relative to that of normal lungs, was enriched with intermediary cells in lung alveolar differentiation with potential to transition to *KRAS*-mutant cancer cells. A subset of alveolar intermediate cells with elevated *KRT8* expression (*KRT8*$^{+}$ alveolar cells; KACs) showed increased plasticity, and their gene expression profiles were enriched in lung precancer and LUAD and signified poor survival. Murine KACs were evident in lungs of tobacco carcinogen-exposed mice that develop *KRAS*-mutant LUADs but not in the saline-treated control group. While murine KACs emerged prior to tumor onset, they persisted for months after carcinogen cessation, acquired driver *Kras* mutations and, like their human counterparts, were poorly differentiated and harbored *KRAS*-specific transcriptional programs. This study provides new insights into early LUAD development and identifies potential targets for treatment.

47 **Main**

48

49 Lung adenocarcinomas (LUAD) represent the most common subtype of lung cancer. Thanks to
50 enhanced screening by low dose CT, more LUADs are being detected at early pathological
51 stages[2-4]. Yet, patient prognosis remains moderate to poor thus posing a growing public health
52 burden that warrants improved early treatment strategies. Decoding the earliest events driving
53 LUADs can inform of ideal targets for its interception. Previous work showed that smoking leads
54 to a pervasive molecular (e.g., *KRAS* mutations) and immune changes that are shared between
55 LUADs and their adjacent normal-appearing ecosystems and that are strongly associated with
56 development of lung premalignant lesions (PMLs) and LUAD[1,5-12]. These earlier studies have
57 shed light on events that are likely implicated in early development of lung tumors driven by
58 somatic mutations in a diverse set of oncogenes (e.g. *KRAS*, *EGFR*, *MET*). Yet, most of these
59 reports were based on bulk approaches and were designed to focus on tumor and distant normal
60 sites in the lung, thereby discounting the cellular and transcriptional phenotypes of the expanded
61 LUAD landscape. Furthermore, while many lung single-cell studies provide details on immune
62 and stromal states[13,14], little insight is drawn to epithelial cells given their paucity (~4%) when
63 performing unbiased scRNA-seq analysis (i.e., without prior enrichment). Therefore, little is
64 known of the identities of specific epithelial subsets or how they promote a field of injury,
65 trigger progression of normal lung to PML, and inspire LUAD pathogenesis. Understanding cell
66 type-specific changes at the root of LUAD inception will help delineate tumor heterogeneity and
67 identify actionable targets and strategies for prevention of this morbid disease. Here, our efforts
68 were focused on in-depth single-cell interrogation of normal and malignant epithelial cells from
69 early-stage LUADs and from an experimental mouse model that faithfully recapitulates the

70 disease, with attention to how specific populations evolve to give rise to malignant tumors.

71

72 **Epithelial transcriptional landscape**

73 Our study design converges in-depth single-cell RNA-sequencing (scRNA-seq) of early-stage
74 LUADs clinical specimens, and, for validation, interrogation of murine lungs from a human-
75 relevant mouse model of LUAD development following exposure to tobacco carcinogen (**Fig.**
76 **1a**). In the human cohort, we studied by scRNA-seq enriched (by sorting for EPCAM+)
77 epithelial cell subsets from early-stage LUADs of 16 patients and 47 paired NL samples (**Fig.**
78 **1a**). As previously described[15], NL tissues were collected at defined spatial proximities from the
79 LUADs – namely, tumor-adjacent, -intermediate, and -distant locations - such that the samples
80 span a topographic continuum (**Fig. 1a**). Our design was purposefully outlined to interrogate
81 normal-appearing lung tissues with variable proximities from the tumors in order to determine
82 cell subsets that are tightly linked with spatial location of the primary tumor. Based on our prior
83 experience as well as others' showing a relatively small fraction of epithelial cells (~4%) in
84 unbiased approaches[15,16], we chose to enrich for epithelial cells, including non-malignant cells in
85 tumor tissues, prior to scRNA-seq. In addition, we collected, from the same tumor and normal
86 regions, a set of tissues for genomic profiling using whole exome sequencing (WES), and
87 another set for high-resolution imaging analysis (**Fig. 1a**).

88

89 Following quality control, we retained 246,102 high-quality epithelial cells for analysis
90 (**Extended Data Fig. 1a, b, Supplementary Table 1**). We identified malignant cells (n =
91 17,064) from otherwise non-malignant "normal" cells (n = 229,038) by integrating information
92 from inferred copy number variations (inferCNV[17]), clustering distribution, lineage-specific gene

93 expression, and the presence of reads carrying *KRAS*^{G12D} somatic mutations (**Fig. 1b-e,**
94 **Extended Data Fig. 1c**). We did not detect any malignant cells from P12 or P16, possibly due to
95 low number of epithelial cells sequenced in these LUADs (Supplementary Data Table 1).
96 Normal epithelial cells we studied also include those that are within the local vicinity of the
97 tumor, i.e., from the LUAD tissues themselves. We identified two main clusters of normal cells,
98 namely alveolar and airway, as well as a small cluster of proliferative cells (**Fig. 1b,**
99 **Supplementary Table 2**). Normal airway epithelial cells (n = 40,607) included diverse lineages
100 such as basal (*KRT17*$^{+}$), ciliated (*FOXJ1*$^{+}$), and club and secretory (*SCGB1A1*$^{+}$) populations, as
101 well as rare types such as ionocytes (*ASCL3*$^{+}$), neuroendocrine (*ASCL1*$^{+}$), and tuft cells
102 (*GNAT3*$^{+}$). Hierarchical clustering showed that club and basal cells were more closely
103 associated, pointing towards overall transcriptomic similarities, in contrast to tuft and ciliated
104 cells which clustered at further distances from the rest of airway cells (**Fig. 1c**). Alveolar cells (n
105 =187,768) consisted of alveolar type 1 (AT1; *AGER1*$^{+}$, *ETV5*$^{+}$), alveolar type 2 (AT2; *SFTPB*$^{+}$,
106 *SFTPC*$^{+}$), bronchioalveolar (BA) that are *SFTPC/SFTPB*+ but with club cell markers, as well as
107 a cluster of alveolar intermediate cells (AICs) that was closely tucked between AT1 and AT2
108 clusters and that shared gene expression features with both major types of alveolar cells (**Fig. 1b,**
109 **c**). Relative to AT1 and AT2 cells, AICs and BA cells were more closely associated with each
110 other - in line with their position on the UMAP and their lineage-intermediate expression
111 features (**Fig. 1c**).
112
113 In contrast to normal cells, malignant cells had low/no expression of lineage-specific markers
114 and showed, overall, loss of lineage identity (**Fig. 1c**). Malignant cells formed 14 clusters (**Fig.**
115 **1d**) that were for the most part patient-specific (**Fig. 1e, left**) signifying strong interpatient

116 heterogeneity. Malignant cells overall showed high levels of aneuploidy based on inferred CNV
117 scores (**Fig. 1e, middle**). We did not detect any unique clustering pattern with respect to smoking
118 status (**Extended Data Fig. 1d**). Annotation based on genomic profiling (by WES) showed that
119 malignant cells from 3 *KRAS*-mutant (KM-) LUAD patients in our cohort (P2, P10, P14)
120 clustered closely together and in comparison to malignant cells from 5 EGFR-mutant, 2 MET-
121 mutant, or other LUADs, which showed a more dispersed clustering pattern (**Fig. 1e, right,**
122 **Extended Data Fig. 1e**). We further confirmed expression of somatic $KRAS^{G12D}$ mutation
123 (inferred from our single-cell data) in malignant cells from the KM-LUAD tumors (**Fig. 1e,**
124 **right**).,
125
126 **LUAD malignant transcriptional programs**
127 We then performed deeper analysis of malignant cells in the context of recurrent driver
128 mutations. Principal component analysis showed that cells from KM-LUADs clustered together
129 and distinctively from cells of *EGFR*-mutant LUADs (EM-LUADs) or *MET*-mutant LUADs
130 (**Fig. 2a**). To further interrogate tumors driven by the same oncogenic drivers, we applied
131 Harmony[18] to adjust for patient-specific effects (see Methods). Among the 7 distinct malignant
132 cell sub-clusters, cluster 5 was primarily enriched with malignant cells from KM-LUAD patients
133 (P2, P10 and P14), (**Extended Data Fig. 1e, Extended Data Fig. 2a**). The vast majority of the
134 *KRAS*-mutant malignant cells clustered together in dimension reduction space and separated
135 from the rest of malignant cells, indicating unique transcriptional reprogramming in *KRAS*-
136 mutant cells (**Fig. 2b**). In line with previous reports[15] showing low levels of chromosomal
137 instability in LUADs driven by strong driver mutations *in vivo*[19], malignant cells from KM-
138 LUADs were chromosomally more stable (indicated by significantly lower CNV scores) in

139 contrast to those from EM-LUADs - which exhibited highest levels of aneuploidy (**Fig. 2b,**
140 **middle**, $p < 0.001$). CNV events were significantly more prominent in malignant cells from
141 smoker relative to never-smoker patients (**Fig. 2b, right**, $p < 0.001$). We noted high degree of
142 inter-patient heterogeneity in the differentiation states of malignant cells, whereby, irrespective
143 of tumor mutation load, cells from KM-LUADs were the least differentiated as indicated by their
144 highest CytoTRACE[20] scores, followed by EM-LUADs (**Fig. 2c**, $p < 0.001$). We also noted
145 intra-tumor heterogeneity (ITH) with respect to differentiation states (e.g. P2, P9, P14, P15),
146 whereby malignant cells from 7 out of the 14 patients with detectable malignant cells (50%)
147 exhibited a broad distribution of CytoTRACE scores, independent of oncogenic mutations (**Fig.**
148 **2c, right**). These results point towards extensive heterogeneity within and across tumors, with
149 malignant cells from KM-LUADs exhibiting distinctively reduced differentiation.
150
151 To better understand malignant cell transcriptional programs that are shared across tumors and
152 govern tumor cell heterogeneity, we measured, in malignant cells from our cohort, the expression
153 levels of 23 recurrent malignant meta-programs (MPs) that were recently defined in the
154 comprehensive study by Gavish and colleagues[21] and which interrogated >1,000 human tumor
155 samples across a large number of scRNA-seq datasets. Unsupervised clustering identified 5 MP-
156 driven cell clusters (Meta_C1- Meta_C5) based on expression [21] (**Fig. 2d, Supplementary Table**
157 **3**). Notably, 96.8% of malignant cells from KM-LUADs were clustered in Meta_C1 which
158 showed distinct expression of these MPs compared to malignant cells in other clusters (**Fig. 2d,**
159 **e, Extended Data Fig. 2b**), in line with these cells exhibiting unique transcriptome profiles (**Fig.**
160 **2a, b**). Consistently, cells from Meta_C1 exhibited the highest expression of a gene module
161 associated with $KRAS^{G12D}$ present in pancreatic ductal adenocarcinoma (MP30)[21], epithelial to

162 mesenchymal transition (EMT-III; MP14) and epithelial senescence (MP19), lowest levels of
163 alveolar differentiation (MP31), in line with the findings above (**Fig. 2c**), overall increased
164 cellular respiration (MP21) particularly in cells with reduced differentiation (**Fig. 2d, f,**
165 **Extended Data Fig. 2c**). Meta_C2 was predominantly composed of malignant cells from EM-
166 LUADs (**Fig. 2d, e**). Analysis of Meta_C2 malignant cells pointed towards MP patterns unique
167 to highly differentiated EM-LUADs such as high expression of alveolar MP31, and the lowest
168 expression of EMT-II (MP13), a glioma-pertinent mesenchymal signature MES (MP16), and
169 hypoxia (MP6) relative to other MP clusters (**Fig. 2d-f, Extended Data Fig. 2c**). In contrast to
170 Meta_C2, Meta_C5 had the highest signatures scores for hypoxia (MP6) and the glioma-
171 pertinent mesenchymal signature (MES, MP16), as well as cell cycle signature enriched with
172 chromatin regulator genes (MP4) (**Fig. 2d, f, Extended Data Fig. 2c**). Notably, Meta_C5 was
173 enriched with the most differentiated malignant cells and almost entirely derived from LUADs
174 from never-smokers and that are neither KM- nor EM-LUADs (**Fig. 2d-f, Extended Data Fig.**
175 **2b, c**). MP analysis uncovered a high degree of malignant transcriptional heterogeneity at the
176 single cell level.
177
178 Given the stark contrast of transcriptional programs in KM-LUADs relative to other LUADs, we
179 further analyzed malignant cell states and heterogeneity in KM-LUADs. We examined cell
180 differentiation states, expression of MPs, as well as presence of *KRAS* mutations and CNVs in
181 malignant cells from P14 KM-LUAD (**Fig. 2g, left**). We found increased activation of MP30
182 signature in a subset of $KRAS^{G12D}$ cells (**Fig. 2g, middle)**, and similarly, diverse activation
183 patterns of other MPs (e.g. cellular respiration, EMT-III, cell cycle) across the mutant cells
184 **(Extended Data Fig. 2d**), pointing to marked heterogeneity in transcriptional programs among

185 *KRAS*^{G12D} cells. Overall, *KRAS*^{G12D} cells showed reduced differentiation (higher CytoTRACE
186 scores) relative to malignant cells without the driver mutation (**Fig. 2g**). Trajectory analysis
187 showed strong correlation between reduced differentiation (increased CytoTRACE scores) and
188 *KRAS*^{G12D} cells (**Fig. 2g, right**) which is concordant with loss of alveolar differentiation (MP31)
189 in KM-LUADs (**Fig. 2d-f**). To further interrogate malignant cell genomic heterogeneity, we
190 performed unsupervised CNV clustering analysis and identified three clusters with differential
191 CNV profiles (**Extended Data Fig. 2e, left**). *KRAS*^{G12D} cells were enriched in cells harboring
192 late CNV events (e.g. Chr 1p loss, Chr 8 and Chr 12 gains, identified by clustering analysis of
193 inferred CNVs, as previously described[15]), and reduced alveolar signature scores in line with
194 attenuated differentiation (**Extended Data Fig. 2e, f**). These results suggest that signaling by
195 mutant *KRAS* likely drives transcriptional reprogramming events in KM-LUAD evolution.
196
197 To better characterize malignant transcriptional programs unique to early-stage KM-LUADs, we
198 derived a *KRAS* activation signature based on unique expression patterns of *KRAS*-mutant
199 malignant cells from our cohort (C5; **Extended Data Fig. 2a**). This signature correlated strongly
200 and significantly with *KRAS* MP30 present in *KRAS*-mutant pancreatic ductal adenocarcinomas
201 (**Extended Data Fig. 2g**; $p < 0.0001$). Further analysis of TCGA LUAD cohort showed that
202 KM-LUADs with relatively high expression of our *KRAS* signature were enriched with activated
203 *KRAS* MP30 as well as other MPs we had found to be increased in Meta_C1 (e.g. epithelial
204 senescence MP19, EMT-III MP14) (n = 498; **Fig. 2h, left**). Relative to *KRAS* signature-low
205 LUADs, LUADs with high expression of the signature displayed significantly reduced overall
206 survival (OS) irrespective of *KRAS* mutation status (**Extended Data Fig. 2h**, **left**; $p = 0.005$). A
207 similar survival disadvantage was observed in *KRAS* signature-high tumors with wild-type

208 *KRAS*, and in *KRAS*^{G12D}-LUADs despite the small cohort size (**Fig. 2i**; $p = 0.02$). Notably, and
209 compared to other more frequently observed *KRAS* codon 12 driver mutations (e.g. *KRAS*^{G12C},
210 *KRAS*^{G12V}), *KRAS*^{G12D}-LUADs clustered closer to KM-LUADs with high *KRAS* signature
211 expression and showed increased activation of associated MPs: *KRAS* (MP30), epithelial
212 senescence (MP19), and EMT-III (MP14) (**Extended Data Fig. 2h, right**; $p < 0.05$). Taken
213 together, our analysis of a large number of malignant cells from early-stage LUADs identified
214 neoplastic transcriptomic heterogeneity, cell-level ITH within a single LUAD region, and a
215 malignant signature that is biologically- and possibly- clinically-relevant to KM-LUADs.
216
217 **Alveolar intermediate cells in LUADs**
218 Our data suggesting epithelial cell plasticity and heterogeneity in early LUADs, prompted us to
219 investigate specific cell subsets and states that are implicated in LUAD development. We
220 analyzed proportions of epithelial lineages (e.g. alveolar and airway) among epithelial cells in
221 LUAD tissues and the multi-region uninvolved normal lung (NL) samples and found no
222 significant changes in fractions of AT1, basal, ciliated, or club and secretory cell populations
223 across the different tissues (**Extended Data Fig. 3a, b**). Notably, the fraction of AT2 cells was
224 significantly reduced with closer proximity to the LUADs ($p = 0.002$) (**Fig. 3a, b**). Interestingly,
225 we noted a distinctively significant enrichment of alveolar intermediate cell (AICs) in the LUAD
226 tissues relative to all NLs ($p = 0.02$) (**Fig. 3a, b**) possibly signifying alveolar changes that are
227 pertinent to the local niche of LUADs. These observations were overall consistent across
228 individual patient samples (**Extended Data Fig. 3a**). Recent reports have identified an alveolar
229 intermediate state which emerges in response to acute lung injury in mice and along the AT2-to-
230 AT1 differentiation axis, a process which serves to replenish damaged lung tissues[22]. Yet, the

231 role of these repair mechanisms in the context of LUAD development in humans are poorly
232 understood. We interrogated developmental and differentiation trajectories among alveolar cells
233 in our cohort and found that AICs represented an intermediary state (intermediate pseudotime
234 and CytoTRACE scores) along the differentiation trajectory of AT2 cells to the relatively highly
235 differentiated AT1 cells (lowest CytoTRACE score) (**Fig. 3c, d**). The proportion of least
236 differentiated AICs in LUAD tissues was higher than that of their more differentiated
237 counterparts (low CytoTRACE scores) (29% versus 11%, respectively, **Extended Data Fig. 3c**).
238 Notably, trajectory analysis suggests that AICs could also transition to malignant cells (**Fig. 3c,**
239 **top branch of the trajectory**). A fraction of these cells harbored driver *KRAS* mutations, which
240 were overall significantly more developmentally late relative to *EGFR*-mutant malignant cells
241 (**Fig. 3d, bottom**). This implies that AICs could be associated with development of LUADs in
242 these patients. Next, we took a deeper dive into AICs and identified two transcriptionally distinct
243 sub-populations whereby one subset featured high expression of *KRT8* (**Fig. 3e**). We refer to
244 these AICs as "*KRT8*-expressing Alveolar Cells" or KACs and found that they had uniquely
245 increased expression of *CDKN1A/2A, PLAUR, CLDN4, LCN2, UC5B* and *TTF3* relative to other
246 AICs (**Fig. 3e, Extended Data Fig. 3d, Supplementary Table 4**). KACs were also significantly
247 less differentiated (higher CytoTRACE score) relative to other AICs (**Fig. 3f**). The proportion of
248 KACs among non-malignant epithelial cells detected in LUAD samples was strongly and
249 significantly increased relative to that among non-malignant epithelial cells of multi-region NL
250 tissues (**Fig. 3g**). Furthermore, the fraction of KACs in LUAD samples was significantly higher
251 than that of AT1, AT2, or other AICs (**Fig. 3h, right**). Of note, KACs were also evident in NL
252 samples and they clustered closer to other AICs relative to tumor derived KACs (**Fig. 3h, left**).
253

254 High-resolution, multi-plex, and spatial imaging analysis of KRT8, CLDN4, and pan-cytokeratin
255 (PanCK) validated the presence of KACs *in situ* in LUAD tissues from multiple patients (**Fig. 3i,**
256 **Extended Data Fig. 3e**). KACs were not only enriched in tumor-adjacent regions (TANs) but
257 were found right next to malignant cells, the latter which highly expressed KRT8 and CLDN4
258 (**Fig. 3i, Extended Data Fig. 3e**). While KACs were also found in the uninvolved NLs,
259 consistent with our scRNAseq-analysis, only in the TANs did they display features of "reactive"
260 or hyperplastic epithelial cells (**Fig. 3i, Extended Data Fig. 3e**). Intriguingly, we found that
261 expression of *KRAS* signature (see **Fig. 2**) was the highest in KACs relative to AT1, AT2, or
262 other AICs, and in stark contrast to alveolar signature (MP31) and AT1 signature (**Fig. 3j, left,**
263 **Extended Data Fig. 3f**). We derived a KAC signature and found that it significantly and
264 positively correlated with *KRAS* signature (r = 0.45; $p < 2.2 \times 10^{-16}$) and inversely correlated with
265 alveolar signature (r = -0.77; $p < 2.2 \times 10^{-16}$; **Fig. 3j, right**). KAC signature was also enriched in
266 malignant cells from KM-LUADs relative to LUADs with wild-type *KRAS* ($p < 0.001$;
267 **Extended Data Fig. 3g**). We then found that the KAC signature was significantly enriched in
268 bulk transcriptomes of human TCGA LUADs compared to uninvolved NL tissues (**Fig. 3k, top**;
269 $p = 1.9 \times 10^{-15}$), and in contrast to alveolar signature which was significantly reduced in LUADs
270 ($p < 0.001$; **Extended Data Fig. 3h**). By analysis of our independent cohort of preneoplasia, we
271 also found that the KAC signature was significantly and progressively increased along the
272 pathologic continuum of NL, atypical adenomatous hyperplasia (AAH; the earliest precursor of
273 LUAD), and matching invasive LUAD (**Fig. 3k, bottom**; $p < 0.05$). The fraction of KM-LUADs
274 among TCGA samples with relatively high KAC signature score was significantly higher than
275 those with low expression of the signature (n = 498, $p < 0.05$, **Fig. 3l, left**). Also, patients with
276 KAC signature-high LUADs (n = 243) showed significantly reduced OS ($p = 0.005$) relative to

277 those with low expression of the signature (n = 246; **Fig. 3l, right**). In a separate cohort, KAC
278 signature score significantly and positively correlated with expression of *LCN2*, a gene we
279 previously found to be strongly associated with KM-LUAD (n = 150; r = 0.6; $p < 10^{-15}$)[23], and
280 was associated with poor OS ($p = 0.04$; **Fig. 3m**). These data describe an intermediate alveolar
281 cell state that is linked to LUAD pathogenesis in humans.
282
283 **A KAC state is linked to murine KM-LUADs**
284 Since the association of KACs with human LUAD was not previously reported and their
285 implication in lung tumor development is not known, we thus sought to validate the association
286 of these cells with LUAD development *in vivo*, and if confirmed, determine their temporal
287 dynamics in pathogenesis of the malignancy. We performed scRNA-seq analysis of lung
288 epithelial cells from mice with knockout of the airway lineage-specific G-protein coupled
289 receptor (*Gprc5a*$^{-/-}$)[24,25] in the presence or absence of tobacco-carcinogen exposure. LUADs
290 form in *Gprc5a*$^{-/-}$ mice but not in their wild-type littermates, are accelerated by tobacco-
291 carcinogen exposure, and acquired somatic *Kras*^{G12D} mutations – we surmised that these features
292 are highly pertinent to KM-LUAD development[24,26,27] and thus to exploring KACs in this setting.
293 We studied by scRNA-seq lungs from *Gprc5a*$^{-/-}$ mice treated with nicotine-derived nitrosamine
294 ketone (NNK) or saline at the end of exposure (EOE) and at 7 months post-exposure, the time
295 point of KM-LUAD onset in all mice (n = 4 mice per group and time point) (**Fig 4a**). Clustering
296 analysis of 9,272 high-quality epithelial cells revealed distinct lineages that were similarly
297 observed in the human lungs including KACs that clustered in between AT1 and AT2 cell
298 subsets and close to tumor cells (**Fig. 4b, c**). Like their human counterparts, malignant cells
299 displayed low expression of lineage-specific genes (**Fig. 4c**). Cells from the malignant cluster

300 harbored high CNV scores, expressed *Kras*^{G12D} mutation, and showed increased expression of
301 markers associated with loss of alveolar differentiation (*Kng2* and *Meg3*) and the epithelial-
302 specific immune checkpoint *Cd24a* which we had previously shown to be associated with LUAD
303 pathogenesis[28] (**Fig. 4d, Extended Data Fig. 4a**). Intriguingly, relative to AT1 and AT2 cells,
304 KACs displayed markedly increased prevalence of *Kras*^{G12D} mutation, more so than CNV
305 burden, and increased expression of gastric differentiation gene *Gnk2*- in line with previous
306 reports[29] and suggestive of reduced alveolar differentiation - albeit to lesser extents compared to
307 malignant cells (**Fig. 4e, Extended Data Fig. 4b**). We then probed temporal dynamics of
308 alveolar cell subsets and malignant cells. Malignant cells were present only at 7 months post-
309 NNK and were absent from samples collected at EOE to carcinogen or in saline-treated animals
310 (**Fig. 4f, g**). In contrast, fractions of KACs were markedly increased early on in NNK-exposed
311 mice relative to control littermates treated with saline at EOE to the carcinogen and prior to
312 tumor formation and, were for the most part, maintained at 7 months post exposure ($p < 0.01$;
313 **Fig. 4f, g; Extended Data Fig. 4c, d**). Of note, fractions of AT2 cells were reduced with time
314 (**Extended Data Fig. 4d, e**) in line with our observations in human LUADs (**Fig. 3b**).
315 Immunofluorescence (IF) analysis of Krt8, Lamp3 (AT2), and Pdpn (AT1) expression in mouse
316 lung tissues showed that KACs were present in NNK-exposed normal lungs and were nearly
317 absent in uninvolved normal lung regions of mice exposed to saline (**Fig. 4h**). Of note, LUADs
318 themselves also displayed prominent expression of Krt8 (**Fig. 4h**).
319
320 Developmental trajectory analysis revealed that while KACs are likely AT2-derived and act as
321 intermediate states in AT2 to AT1 differentiation (from left to right), in close agreement with
322 recent reports in the context of acute lung injury[30,31] and our analysis of human LUADs (**Fig. 3**),

323 they also likely transitioned to malignant cells (**Fig. 5a, top row**). KACs assumed an
324 intermediate differentiation state (resembling malignant cells) along the path from least
325 differentiated AT2 cells and up to the most differentiated AT1 cells (highest to lowest
326 CytoTRACE scores), thereby suggesting that KACs may well represent tumor progenitors (**Fig.**
327 **5a**). Intriguingly, a closer look showed that KACs at EOE were-less differentiated (higher
328 CytoTRACE scores) relative to the more differentiated KACs at 7 months post-exposure (**Fig.**
329 **5a, bottom right**). Indeed, we found that KACs were enriched in organoids derived from AT2
330 cells of NNK- but not saline-exposed *Gprc5a*$^{-/-}$ mice with reporter labelled-AT2 cells
331 (*Sftpc*$^{CreER/+}$; *Rosa*$^{Sun1GFP/+}$; *Gprc5a*$^{-/-}$) (**Extended Data Fig. 4f**). Analysis of differential gene
332 expression among KACs, AT1, AT2, and malignant cells, showed that KACs clustered most
333 closely with tumor cells (**Fig. 5b**). Very much like their human counterparts (**Fig. 3**), KACs
334 displayed markedly upregulated expression of *Krt8*, *Cldn4*, key players in p53 signaling
335 including *Trp53* itself, and *Cdkn2a,* the negative regulator of cell cycle(**Fig. 5c**). Single-cell
336 pathway analysis showed that NF-κB, p53 signaling, and EMT were the top pathways enriched
337 or activated in KACs (**Fig. 5d**). We derived a murine KAC signature and found that it was
338 gradually and significantly enriched in AT2 cells and up to KACs and malignant cells ($p < 0.05$;
339 **Fig. 5e**). The murine KAC signature was also enriched in human LUADs, as well as gradually
340 increased in transcriptomes of premalignant AAHs and invasive LUADs, and relative to matched
341 NL tissues (**Extended Data Fig. 5g**). The mouse KAC signature partially overlaps (21%) with
342 the human KAC counterpart (**Fig. 3**), and significantly so when compared to overlap by chance
343 (3%, $p < 0.001$; **Fig. 5f**). Expression of human KAC marker gene orthologs was highly enriched
344 in mouse KACs relative to other alveolar cells (**Fig. 5g**). Our *in vivo* analyses identify KACs as

345 an intermediate cell state in early development of KM-LUAD and following tobacco carcinogen
346 exposure.
347
348
349 **Discussion**
350
351 Our high-resolution analysis of epithelial cells from early-stage LUADs and the peripheral lung
352 uncovered diverse malignant states, patterns of ITH, and cell plasticity programs that are linked
353 to KM-LUAD pathogenesis. Our analysis of "normal" epithelial cell subsets from LUADs and
354 the uninvolved normal lungs describe a scenario whereby intermediary cells that arise upon
355 activation of alveolar differentiation programs are linked to transformation of normal lung
356 epithelial cells to malignant states (**Fig. 5h**). AICs can transition to KACs, which themselves
357 likely comprise LUAD progenitors, and emerge early on and acquire unique transcriptional
358 programs, disrupted differentiation, and activation of epithelial plasticity programs - phenotypes
359 that are important for LUAD development. We also show that AICs temporally evolve early on
360 post tobacco carcinogen exposure *in vivo* and prior to tumor onset, thereby lending support to the
361 notion that these cells could be a discrete phenotypic attribute of LUAD *field of injury*. It is
362 plausible that the unique and evolutionarily conserved AIC states are at the root of LUAD
363 inception and therefore, comprise a feature of a primed normal-appearing or premalignant
364 injured/mutated cells[28] (**Fig. 5h**). Thus, our discoveries inspire the derivation of targets to
365 prevent initiation and development of this deadly disease.
366

367 Our study catalogs a rich cellular and transcriptional landscape of lung epithelial cells from
368 early-stage LUAD patients. We describe remarkable inter-patient heterogeneity of malignant cell
369 subsets that is likely based, and possibly partially driven by, different driver mutations. Of note,
370 we uncover unique transcriptomic aspects of *KRAS*-mutant LUADs, whereby loss of epithelial
371 lineage identity was a hallmark of malignant cells from these tumors. Our deep single-cell
372 profiling of *KRAS*-mutant cells was not only reminiscent of previous reports showing loss of
373 AT2 identity in organoid models of KM-LUAD[32], and lineage promiscuity in metastatic
374 LUADs[16], but also pinpointed marked heterogeneity in transcriptional programs among these
375 cells. In-depth interrogation of a large number of malignant cells also enabled us to derive a
376 *KRAS* signature with potential translational applications - this signature marks *KRAS*-mutant
377 tumors when examined in public datasets and is significantly associated with poor survival.
378 These findings point towards possible treatment targets for KM-LUAD that may be contextually
379 relevant to emerging therapeutic strategies that target *KRAS*[33]. Our results also motivate a timely
380 interrogation of potential combinatorial approaches that could amplify the benefits of next-
381 generation $KRAS^{G12D}$ or pan-*KRAS* inhibitors for treatment of this highly disadvantaged
382 population of patients.
383
384 We also had the capacity to study our large cohort of non-malignant epithelial cells in more
385 detail. In human LUADs, we found AICs which, for the most part, had been previously
386 described in animal models as intermediate cells in a transitionary state between AT1 and AT2
387 cells[30,31,34]. This led us to identify tumor-associated KACs, a subset of AICs that arises from AT2
388 cells and that is vastly enriched in LUADs. Our results support the existence of alveolar
389 plasticity programs that are active early on in LUAD development. Recent studies have shown

390 that following acute lung injury, AT2 cells differentiate into AICs that are characterized by high
391 expression of *Krt8* and that are crucial for regeneration of AT1 cells[30,31,34]. In parallel, few
392 studies have described cellular disruptive contexts whereby regeneration programs constitute
393 mechanisms that not only could underlie cancer development, but that may also drive malignant
394 cell heterogeneity[14,35]. Despite these findings, the implications of AT2-mediated repair
395 mechanisms in lung tumorigenesis have been for the most part, largely elusive. We propose that
396 KACs could short-circuit the AT2-to-AT1 differentiation trajectory, thereby transitioning to
397 malignant cells including those harboring driver *KRAS*^{G12D} mutations (**Fig. 5h**). We also
398 characterized a KAC signature that associated with poor cell differentiation status and
399 augmented epithelial plasticity, a functionally-discrete hallmark of cancer[36]. Congruent with
400 these findings, we showed that tobacco carcinogen exposure *in vivo* activated cellular programs
401 that enabled the phenotypic switching of AT2 cells to KACs and which were closely linked to
402 emergence of malignant cells in KM-LUADs, suggesting that the early appearance of these cells
403 following tobacco exposure may represent development of a *field of injury*[11]. Additional studies
404 can further unravel the role of plastic cell states in the continuum of lung malignancy and
405 identify potential KAC signals that could be harnessed as targets for prevention of PML
406 inception and/or progression at the earliest possible stages.
407
408 From our downstream analysis of KACs, we uncovered multiple lines of evidence pointing
409 towards a strong link between KACs and LUADs that specifically harbor *KRAS* mutations. First,
410 we found that KACs, much like *KRAS*-mutant malignant cells, are less differentiated relative to
411 AT1 and AT2 subsets. This is in line with our analysis of malignant cells in KM-LUAD and
412 whereby loss of alveolar differentiation was a unique characteristic of these cells. Furthermore,

413 gene expression in both KACs and malignant *KRAS* mutant tumors showed activation of EMT
414 and mutant p53 programs. Intrigungly, some of these transcriptional modules were recently
415 reported in mice undergoing lung repair in respose to acute injury[30,31,34]. Furthemore, while our
416 *in vivo* findings are consistent with previous independent reports showing that AT2 cells are the
417 preferential cells-of-origin in *Kras*-driven murine LUADs in aimals[37-39], our results enable a
418 deeper scrutiny of the unique attributes and states of alveolar intermediary cells that have
419 potential to constitute progenitors for KM-LUADs. Future analysis of additional LUAD tissues
420 as well as mouse models of LUAD driven by other oncogenes (such as *EGFR*) can serve to
421 determine whether KACs are specific to KM-LUADs.
422
423 Our study design converges an in-depth interrogation of early-stage LUADs as well as a human-
424 relevant mouse model of KM-LUAD which enabled the analysis of an expansive number of
425 epithelial cells. This atlas of epithelial and malignant cell states in human and murine LUADs
426 underscores new cell-specific states that underlie inception of LUADs, and that can thus guide
427 novel strategies to prevent the initiation and development of this morbid disease.
428
429 **References**

430 1 Kadara, H., Scheet, P., Wistuba, II & Spira, A. E. Early Events in the Molecular
431 Pathogenesis of Lung Cancer. *Cancer Prev Res (Phila)* **9**, 518-527, doi:10.1158/1940-
432 6207.CAPR-15-0400 (2016).
433 2 Cardarella, S. & Johnson, B. E. The impact of genomic changes on treatment of lung
434 cancer. *American journal of respiratory and critical care medicine* **188**, 770-775,
435 doi:10.1164/rccm.201305-0843PP (2013).

436 3 Timar, J. The clinical relevance of KRAS gene mutation in non-small-cell lung cancer.
437 *Current opinion in oncology* **26**, 138-144, doi:10.1097/CCO.0000000000000051 (2014).
438 4 Tomasini, P., Walia, P., Labbe, C., Jao, K. & Leighl, N. B. Targeting the KRAS Pathway
439 in Non-Small Cell Lung Cancer. *The oncologist* **21**, 1450-1460,
440 doi:10.1634/theoncologist.2015-0084 (2016).
441 5 Kadara, H. *et al.* Transcriptomic architecture of the adjacent airway field cancerization in
442 non-small cell lung cancer. *J Natl Cancer Inst* **106**, dju004, doi:10.1093/jnci/dju004
443 (2014).
444 6 Belinsky, S. A. *et al.* Aberrant methylation of p16(INK4a) is an early event in lung
445 cancer and a potential biomarker for early diagnosis. *Proc Natl Acad Sci U S A* **95**,
446 11891-11896 (1998).
447 7 Belinsky, S. A. *et al.* Aberrant promoter methylation in bronchial epithelium and sputum
448 from current and former smokers. *Cancer Res* **62**, 2370-2377 (2002).
449 8 Franklin, W. A. *et al.* Widely dispersed p53 mutation in respiratory epithelium. A novel
450 mechanism for field carcinogenesis. *J Clin Invest* **100**, 2133-2137,
451 doi:10.1172/JCI119748 (1997).
452 9 Spira, A. *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers
453 with suspect lung cancer. *Nat Med* **13**, 361-366, doi:nm1556 [pii] 10.1038/nm1556
454 (2007).
455 10 Spira, A., Halmos, B. & Powell, C. A. Update in Lung Cancer 2014. *American journal of*
456 *respiratory and critical care medicine* **192**, 283-294, doi:10.1164/rccm.201504-0756UP
457 (2015).

458 11 Steiling, K., Ryan, J., Brody, J. S. & Spira, A. The field of tissue injury in the lung and
459 airway. *Cancer Prev Res (Phila Pa)* **1**, 396-403, doi:1/6/396 [pii] 10.1158/1940-
460 6207.CAPR-08-0174 (2008).

461 12 Wistuba, II *et al.* Molecular damage in the bronchial epithelium of current and former
462 smokers. *J Natl Cancer Inst* **89**, 1366-1373 (1997).

463 13 Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor
464 microenvironment. *Nat Med* **24**, 1277-1289, doi:10.1038/s41591-018-0096-5 (2018).

465 14 Maynard, A. *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by
466 Single-Cell RNA Sequencing. *Cell* **182**, 1232-1251.e1222,
467 doi:10.1016/j.cell.2020.07.017 (2020).

468 15 Sinjab, A. *et al.* Resolving the Spatial and Cellular Architecture of Lung
469 Adenocarcinoma by Multiregion Single-Cell Sequencing. *Cancer discovery* **11**, 2506,
470 doi:10.1158/2159-8290.CD-20-1285 (2021).

471 16 Laughney, A. M. *et al.* Regenerative lineages and immune-mediated pruning in lung
472 cancer metastasis. *Nat Med* **26**, 259-269, doi:10.1038/s41591-019-0750-6 (2020).

473 17 Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human
474 oligodendroglioma. *Nature* **539**, 309-313, doi:10.1038/nature20123 (2016).

475 18 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with
476 Harmony. *Nature methods* **16**, 1289-1296, doi:10.1038/s41592-019-0619-0 (2019).

477 19 McFadden, D. G. *et al.* Mutational landscape of EGFR-, MYC-, and Kras-driven
478 genetically engineered mouse models of lung adenocarcinoma. *Proc Natl Acad Sci U S A*
479 **113**, E6409-e6417, doi:10.1073/pnas.1613601113 (2016).

480 20 Gulati, G. S. *et al.* Single-cell transcriptional diversity is a hallmark of developmental
481 potential. *Science* **367**, 405-411, doi:10.1126/science.aax0249 (2020).
482 21 Gavish, A. *et al.* The transcriptional hallmarks of intra-tumor heterogeneity across a
483 thousand tumors. *bioRxiv*, 2021.2012.2019.473368, doi:10.1101/2021.12.19.473368
484 (2021).
485 22 Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using
486 single-cell RNA-seq. *Nature* **509**, 371-375, doi:10.1038/nature13173 (2014).
487 23 Treekitkarnmongkol, W. *et al.* Augmented Lipocalin-2 Is Associated with Chronic
488 Obstructive Pulmonary Disease and Counteracts Lung Adenocarcinoma Development.
489 *American journal of respiratory and critical care medicine* **203**, 90-101,
490 doi:10.1164/rccm.202004-1079OC (2021).
491 24 Fujimoto, J. *et al.* Development of Kras mutant lung adenocarcinoma in mice with
492 knockout of the airway lineage-specific gene Gprc5a. *International journal of cancer*
493 **141**, 1589-1599, doi:10.1002/ijc.30851 (2017).
494 25 Westcott, P. M. *et al.* The mutational landscapes of genetic and chemical models of Kras-
495 driven lung cancer. *Nature* **517**, 489-492, doi:10.1038/nature13898 (2015).
496 26 Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung
497 adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).
498 27 Kantrowitz, J. *et al.* Genome-Wide Gene Expression Changes in the Normal-Appearing
499 Airway during the Evolution of Smoking-Associated Lung Adenocarcinoma. *Cancer*
500 *Prev Res (Phila)* **11**, 237-248, doi:10.1158/1940-6207.CAPR-17-0295 (2018).

501 28 Sinjab, A., Han, G., Wang, L. & Kadara, H. Field Carcinogenesis in Cancer Evolution:
502 What the Cell Is Going On? *Cancer Research* **80**, 4888, doi:10.1158/0008-5472.CAN-
503 20-1956 (2020).
504 29 Snyder, E. L. *et al.* Nkx2-1 represses a latent gastric differentiation program in lung
505 adenocarcinoma. *Mol Cell* **50**, 185-199, doi:10.1016/j.molcel.2013.02.018 (2013).
506 30 Choi, J. *et al.* Inflammatory Signals Induce AT2 Cell-Derived Damage-Associated
507 Transient Progenitors that Mediate Alveolar Regeneration. *Cell Stem Cell* **27**, 366-
508 382.e367, doi:10.1016/j.stem.2020.06.020 (2020).
509 31 Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that
510 persists in human lung fibrosis. *Nature Communications* **11**, 3559, doi:10.1038/s41467-
511 020-17358-3 (2020).
512 32 Dost, A. F. M. *et al.* Organoids Model Transcriptional Hallmarks of Oncogenic KRAS
513 Activation in Lung Epithelial Progenitor Cells. *Cell Stem Cell* **27**, 663-678.e668,
514 doi:10.1016/j.stem.2020.07.022 (2020).
515 33 Salgia, R., Pharaon, R., Mambetsariev, I., Nam, A. & Sattler, M. The improbable targeted
516 therapy: KRAS as an emerging target in non-small cell lung cancer (NSCLC). *Cell*
517 *Reports Medicine* **2**, 100186, doi:https://doi.org/10.1016/j.xcrm.2020.100186 (2021).
518 34 Kobayashi, Y. *et al.* Persistence of a regeneration-associated, transitional alveolar
519 epithelial cell state in pulmonary fibrosis. *Nat Cell Biol* **22**, 934-946,
520 doi:10.1038/s41556-020-0542-8 (2020).
521 35 Kadur Lakshminarasimha Murthy, P. *et al.* Human distal lung maps and lineage
522 hierarchies reveal a bipotent progenitor. *Nature* **604**, 111-119, doi:10.1038/s41586-022-
523 04541-3 (2022).

524 36 Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer discovery* **12**, 31-46,
525 doi:10.1158/2159-8290.Cd-21-1059 (2022).
526 37 Lin, C. *et al.* Alveolar type II cells possess the capability of initiating lung tumor
527 development. *PloS one* **7**, e53817, doi:10.1371/journal.pone.0053817 (2012).
528 38 Liu, B. *et al.* An entropy-based metric for assessing the purity of single cell populations.
529 *Nature Communications* **11**, 3155, doi:10.1038/s41467-020-16904-3 (2020).
530 39 Xu, X. *et al.* Evidence for type II cells as cells of origin of K-Ras-induced distal lung
531 adenocarcinoma. *Proc Natl Acad Sci U S A* **109**, 4910-4915,
532 doi:10.1073/pnas.1112499109 (2012).
533
534
535 **Figure legends**
536
537 **Fig. 1. Transcriptional landscape of lung epithelial cells in early-stage LUAD patients. a,**
538 schematic overview of the experimental design and the analysis workflow. State: cellular
539 transcriptional state. Composition: composition of cell subsets. Program: transcriptional
540 programs in malignant cells. **b,** UMAP plots of 229,038 normal epithelial cells from 63 samples.
541 Each dot represents a single cell colored by major cell lineages (left), airway sub lineages (top)
542 and alveolar sub lineages (bottom). AIC: alveolar intermediate cell. **c,** Sub-lineage hierarchy
543 within airways cells and alveolar cells (left) and expression (proportions and average expression
544 levels) of selected marker genes for 10 normal epithelial cell types as defined in panel b (right)
545 and malignant cells. **d,** Unsupervised clustering of 17,064 malignant cells colored by cluster ID.
546 The zoom-in view (top left) showing malignant cells colored by $KRAS^{G12D}$ mutation status

547 identified using scRNA-seq data. **e,** UMAP plots of malignant cells as shown in **d,** but colored
548 by patient ID (left), inferred CNV score (middle), and driver mutations identified in each tumor
549 sample using whole-exome sequencing (right).
550
551 **Fig. 2. Deep single-cell analysis of malignant cells reveals unique signatures, states, and**
552 **malignant programs. a,** Principal component analysis (PCA) plot of malignant cells colored by
553 driver mutations identified in each tumor sample using whole-exome sequencing. **b,** Harmony-
554 corrected UMAP plots of malignant cells colored by patient ID (left, which were assigned to
555 color palette groups based on driver mutation status), CNV scores (top middle), smoking status
556 (top right). Comparison of CNV scores between malignant cells from samples carrying different
557 driver mutations (bottom left) or between smokers and never smokers (bottom right). ***: FDR
558 adjusted p-value < 0.001. Box, median ± interquartile range; whiskers, 1.5× interquartile range;
559 center line: median. **c,** Landscape of malignant cell differentiation status (top left). Comparison
560 of differentiation states (inferred by CytoTRACE) between malignant cells from samples with
561 different driver mutations (bottom left). Per sample distribution of cell CytoTRACE score
562 (right). TMB: tumor mutation burden. Box definitions are the same as in panel **b**. **d,**
563 Unsupervised clustering of malignant cells based on expression of 23 previously defined
564 consensus intratumor heterogeneity meta-programs (MPs). **e,** Enrichment of clusters (C1-C5)
565 from panel **d** in samples with different driver mutations. Distribution of CNV scores (top right)
566 and CytoTRACE scores (bottom right) across the 5 clusters defined in panel **d**. Box definitions
567 are the same as in panel **b**. **f,** Distribution of signature scores of 4 representative MPS: across
568 clusters defined in panel **d** in panel d. Box definitions are the same as in panel **b**. **g,** Profiling of
569 intratumoral heterogeneity in malignant cells from P14. UMAP plots show P14 malignant cells

570 colored by (left to right) *KRAS*^{G12D} mutation status, *KRAS* signature expression, and cell
571 differentiation status (CytoTRACE), respectively. Developmental trajectory of P14 malignant
572 cells colored by inferred pseudotime, differentiation status, and the presence of *KRAS*^{G12D}
573 mutations. **h,** *KRAS* signature expression in TCGA LUAD cohort. Enrichment of significantly
574 correlated meta-programs with *KRAS* signature in TCGA LUAD cohort are shown. Samples
575 harboring somatic *KRAS*^{G12D} mutations are annotated on the top. **i,** Survival probability
576 differences between samples carrying somatic *KRAS*^{G12D} mutation, and those with high or low
577 level of *KRAS* signature expression. *KRAS* sig. high: samples showing high *KRAS* signature
578 expression (the top quartile) . *KRAS* sig. low: samples with low *KRAS* signature expression
579 (below the third quartile). *KRAS*^{G12D}: samples with *KRAS*^{G12D} mutation.
580
581 **Fig. 3. Identification and characterization of *KRT8*+ alveolar intermediate cells that are**
582 **implicated in human LUAD development. a,** Composition of normal epithelial lineages across
583 spatial regions as defined in **Fig. 1a**. Dis: distant normal. Int: intermediate normal. Adj: adjacent
584 normal. AIC: alveolar intermediate cells. AT1: alveolar type 1. AT2: alveolar type 2. BA:
585 bronchioalveolar. NE: neuroendocrine. **b,** Changes in cellular fractional of AT2 cells (left) and
586 AICs (right) across the spatial samples. *P*-values were calculated with Kruskal-Wallis test. Box,
587 median ± interquartile range; whiskers, 1.5× interquartile range; center line: median. **c,**
588 Pseudotime trajectory analysis of alveolar and malignant cells identified in this study. **d,**
589 Distribution of CytoTRACE scores in AICs, AT1 and AT2 cells (top). Distribution of
590 pseudotime scores in malignant cells of *EGFR*- and *KRAS*-mutant tumors. ***: p-value < 0.001.
591 *P*-value was calculated with Mann-Whitney U test. Box definitions are the same as in panel **b. e,**
592 Sub-clustering analysis of AICs (left). KACs: *KRT8*$^{+}$ alveolar intermediate cells. Expression

593 levels and frequencies of representative marker genes of KACs (right). **f,** CytoTRACE score in
594 KACs versus other AICs. Box definitions are the same as in panel **b. g,** Proportion of KACs
595 among non-malignant epithelial cells across the spatial samples. *P*-value was calculated using
596 Kruskal-Wallis test. Box definitions are the same as in panel **b. h,** Differences in cell density
597 between LUAD (left) and normal lung tissues (middle). NL: normal lung. Barplot (right) shows
598 fraction of alveolar cell subsets colored by sample type. ***: FDR adjusted *p*-value < 0.001. *P*-
599 values were calculated using Fisher's exact tests. **i,** Microphotographs of P14 LUAD (1-5) and
600 paired uninvolved NL tissue (6-10). Top panels: Hematoxylin & Eosin staining showing LUAD
601 tumor area (T) and tumor adjacent normal tissue (TAN) displaying reactive hyperplasia of AT2
602 cells and presence of intra-alveolar macrophages (1, top left), while 6 shows TAN.
603 Immunofluorescence staining with KRT8 (2 and 7, red), CLDN4 (3 and 8, yellow), PanCK (4
604 and 9, green), and composite image of all three antibodies (5 and 10). Blue nuclear staining was
605 done using Syto13. Magnification 20x. Bar 200 μm. **j,** Enrichment of *KRAS* signature score (top
606 left) and the alveolar lineage signature score (bottom left) across different subsets of alveolar
607 cells. Correlation analysis between KAC signature scores and *KRAS* signature scores (top right)
608 or alveolar lineage signature scores (bottom right). *P*-values were calculated with Spearman
609 correlation test. R denotes the Spearman correlation coefficients. Box definitions are the same as
610 in panel **b. k,** Enrichment of KAC signature score across samples of TCGA LUAD (top) and
611 premalignancy (bottom) cohorts. AAH: atypical adenomatous hyperplasia. *P*-values were
612 calculated using Wilcoxon signed-rank test. Box definitions are the same as in panel **b. l,**
613 Fraction of *KRAS*-mutant samples among TCGA samples with high or low KAC signature score
614 (left). MUT: *KRAS* mutant. WT: *KRAS* wild type. Low: LUAD samples with signature scores
615 lower than the group median value. High: LUAD samples with signature scores higher than the

616 group median value. Differences in overall survival probability between samples with high
617 *versus* low KAC signature scores (right). Mo: months. **m,** Correlation between the expression
618 levels of *LCN2* and KAC signature scores in the PROSPECT cohort (left). Overall survival
619 probability differences between samples with high *versus* low KAC signature scores in
620 PROSPECT cohort (right). Group median value of KAC signature scores was used dichotomize
621 the high and low groups.
622
623 **Fig. 4. Krt8+ alveolar intermediate cells evolve early and prior to tumor onset *in vivo***
624 **during tobacco-associated KM-LUAD pathogenesis. a,** Schematic view of the experimental
625 design. NNK: nicotine-derived nitrosamine ketone. **b,** UMAP view of mouse epithelial cell
626 subsets. **c,** Proportions and average expression levels of selected marker genes for mouse normal
627 epithelial cell lineages and malignant cells as defined in panel **a**. NE: neuroendocrine. KACs:
628 *Krt8+* alveolar cells. **d,** UMAP distribution of alveolar and malignant cells colored by CNV
629 score (top left), $Kras^{G12D}$ mutation status (top right), and expression level of oncogenic markers
630 *Kng2* and *Meg3*. **e,** Distribution of CNV scores (left) among alveolar and malignant cells and
631 fraction of $Kras^{G12D}$ mutant cells in KACs, malignant, AT1 and AT2 subsets (right). Absolute
632 numbers of $Kras^{G12D}$ mutant cells are indicated on top of each bar. **f,** Enlarged UMAPs of
633 alveolar and malignant cells colored by cell subsets (left), $Kras^{G12D}$ mutation status (middle) and
634 CNV level (right), both at end-of-exposure to NNK (top row) and at 7 months post-NNK
635 exposure (bottom row). EOE: end of exposure. Mo: months. **g,** Fraction of malignant cells (left)
636 and KACs (right) across treatment groups and time points. *: FDR adjusted p-value < 0.05. P-
637 values were calculated using Mann-Whitney U test. Box, median ± interquartile range; whiskers,

638 1.5× interquartile range; center line: median. **h,** Immunofluorescence of Krt8, Lamp3 and Pdpn
639 level in mouse lung tissues. NL: normal lung. Scale bars: 10 μm.
640
641 **Fig. 5. *Krt8*+ alveolar cells are implicated in transition of alveolar cells to *Kras*-mutant**
642 **malignant cells. a,** Developmental trajectory of alveolar and malignant cells colored by inferred
643 pseudotime (top left), cell differentiation status (top middle), and cell type (top right).
644 Distribution of inferred pseudotime (bottom left) and CytoTRACE (bottom middle) across the
645 indicated cell subsets. CytoTRACE score distribution in KACs at the two time points. EOE: end
646 of exposure (bottom right). 7 mo: 7 months post exposure to NNK. Box, median ± interquartile
647 range; whiskers, 1.5× interquartile range; center line: median. **b,** Normalized expression level of
648 differentially expressed genes in AT1, KACs, malignant cells and AT2 cells. **c,** Expression levels
649 of representative KAC marker genes across 4 indicated cell subsets. Box, median ± interquartile
650 range; whiskers, 1.5× interquartile range; center line: median. **d,** Significantly enriched
651 functional pathways of marker genes of KACs. **e,** Enrichment of KAC signature score, defined
652 using marker genes of KACs, across the four cell subsets. ***: FDR q-value < 0.001; *: FDR q-
653 value < 0.05. *P*-values were calculated using Mann-Whitney U test. **f,** Fraction of genes
654 belonging to human KAC signature among mouse KAC genes *versus* in genome background. *P*-
655 value was calculated using Fisher's exact test. **g,** Proportions and average expression levels of
656 mouse orthologs of selected human KAC marker genes in murine KACs and other alveolar cells.
657 **h,** Proposed model for alveolar plasticity dynamics in lung cancer whereby a subset of the AICs
658 in the intermediate AT2-to-AT1 differentiation state can evolve into KACs and later, lead to
659 LUAD development.
660

661

662 **METHODS**

663

664 **Multiregional sampling of surgically resected LUADs and normal lung tissues**

665 Study subjects were evaluated at MD Anderson Cancer Center and underwent standard-of-care
666 surgical resection of early-stage LUAD (I–IIIA). All samples in the study were obtained under
667 waiver of consent from banked or residual tissues approved by MD Anderson institutional
668 review board protocols. Residual surgical specimens were then used for derivation of
669 multiregional samples for single-cell analysis (**Supplementary Table 1**). Immediately following
670 surgery, resected tissues were processed by an experienced pathologist assistant. One side of the
671 specimen was documented and measured, followed by tumor margin identification. On the basis
672 of the placement of the tumor within the specimen, incisions were made at defined collection
673 sites in one direction along the length of the specimen and spanning the entire lobe: tumor-
674 adjacent and tumor-distant normal parenchyma at 0.5 cm from the tumor edge and from the
675 periphery of the overall specimen/lobe, respectively. An additional tumor-intermediate normal
676 tissue was selected for P2–P16 and ranged between 3 and 5 cm from the edge of the tumor.
677 Sample collection was initiated at normal lung tissues that are farthest from the tumor moving
678 inward toward the tumor to minimize cross-contamination during collection.

679

680 **Single-cell isolation from LUADs and normal spatial samples**

681 Fresh tissues were collected in RPMI medium supplemented with 2% fetal bovine serum (FBS)
682 and maintained on ice for immediate processing. Tissues were placed in a cell culture dish
683 containing Hank's balanced salt solution (HBSS) on ice and extra-pulmonary airways and

684 connective tissue were removed with scissors. Samples were transferred to a new dish on ice and
685 minced into tiny pieces (approximately 1mm^3) followed by enzymatic digestion using a cocktail
686 containing Collagenase A (10103578001, Sigma Aldrich), Collagenase IV (NC9836075, Thermo
687 Fisher Scientific), DNase I, (11284932001, Sigma Aldrich), Dispase II (4942078001, Sigma
688 Aldrich), Elastase 1 (NC9301601, Thermo Fisher Scientific) and Pronase (10165921001, Sigma
689 Aldrich), as previously described [40]. Samples were incubated in a 37 °C oven for 10 minutes
690 with gentle rotation, pipet mixed at room temperature for further disaggregation, and then
691 incubated at 37 °C for an additional 10 minutes. Samples were then filtered through 70 µm
692 strainers (Miltenyi biotech, 130-098-462) and washed with ice-cold HBSS. Filtrates were then
693 centrifuged and re-suspended in ice-cold ACK lysis buffer (A1049201, Thermo Fisher
694 Scientific) for red blood cell (RBC) lysis. Following RBC lysis, samples were centrifuged and
695 re-suspended in ice-cold FBS, filtered (using 40 µm FlowMi tip filters; H13680-0040,
696 Millipore), and an aliquot was taken to enumerate cells and check for viability by Trypan blue
697 exclusion analysis.
698
699 **Sorting and enrichment of viable lung EPCAM+ epithelial singlets**
700 Single cells from patient 1 (P1) were stained with SYTOX Blue viability dye (S34857, Life
701 Technologies), and processed on a fluorescence-activated cell sorting (FACS) Aria I instrument.
702 Cells from patients 2 through 16 (P2-P16) were stained with anti-EPCAM-PE (347198, BD
703 Biosciences;1:50 dilution in ice-cold phosphate-buffered saline, PBS, containing 2% FBS) for 30
704 minutes with gentle rotation at 4°C. EPCAM-stained cells were then washed, filtered using 40
705 µm tip filters, stained with SYTOX Blue and processed on a FACS Aria II instrument. Doublets
706 and dead cells were eliminated, and f viable (SYTOX-negative) EPCAM-positive (EPCAM+,

707 epithelial) singlets were collected in PBS containing 2% FBS. Cells were washed again to
708 eliminate ambient RNA, and a sample was taken for counting by Trypan Blue exclusion (T8154,
709 Sigma Aldrich) before loading on 10X Chromium microfluidic chips.
710
711 **Preparation of single-cell 5' gene expression libraries**
712 Up to 10,000 cells per sample were partitioned into nanoliter-scale Gel beads-in-emulsion
713 (GEMs) using Chromium Next GEM Single Cell 5' Gel Bead Kit v1.1 (1000169, 10XGenomics)
714 and by loading onto Chromium Next GEM Chips G (1000127, 10XGenomics). GEMs were then
715 recovered to construct single-cell 5' gene expression libraries using the Chromium Next GEM
716 Single Cell 5' Library kit (1000166, 10XGenomics) according to the manufacturer's protocol.
717 Briefly, recovered barcoded GEMs were broken and pooled, followed by magnetic bead clean-up
718 (Dynabeads MyOne Silane, 37002D, Thermo Fisher Scientific). 10X-barcoded full-length cDNA
719 was then amplified by PCR and analyzed using Bioanalyzer High Sensitivity DNA kit (5067-
720 4626, Agilent). Up to 50 ng of cDNA was carried over to construct gene expression libraries and
721 was enzymatically fragmented and size-selected to optimize the cDNA amplicon size prior to 5'
722 gene expression library construction. Further, samples were subject to end-repair, A-tailing,
723 adaptor ligation, and sample index PCR using Single Index Kit T Set A (2000240,
724 10XGenomics) to generate Illumina-ready barcoded gene expression libraries. Library quality
725 and yield was measured using Bioanalyzer High Sensitivity DNA (5067-4626, Agilent) and
726 Qubit dsDNA High Sensitivity Assay (Q32854, Thermo Fisher Scientific) kits. Indexed libraries
727 were normalized by adjusting for the ratio of the targeted cells per library as well as individual
728 library concentration and then pooled to a final concentration of 10 nM. Library pools were then
729 denatured and diluted as recommended for sequencing on the Illumina NovaSeq 6000 platform.

730

731 **Single-cell RNA-seq quality control**

732 The raw scRNA-seq data were pre-processed (demultiplex cellular barcodes, read alignment, and
733 generation of gene count matrix) using Cell Ranger Single Cell Software Suite (version 3.0.1)
734 provided by 10X Genomics. For reads alignment of human and mouse scRNA-seq data, the
735 human reference, GRCh38 (hg38) and mouse reference, GRCm38 (mm10) were used. Detailed
736 quality control (QC) metrics were generated and evaluated, samples and cells were carefully and
737 rigorously filtered to obtain high-quality data for downstream analyses[15]. Briefly, for basic
738 quality filtering, cells with low complexity libraries (in which detected transcripts are aligned to
739 < 200 genes) were filtered out and excluded from subsequent analyses. This filter will help
740 remove cell debris, empty drops, and low-quality cells. Likely dying or apoptotic cells where >
741 15% of transcripts derived from the mitochondrial genome were also excluded.

742

743 **Doublets detection and removal and batch effect evaluation and correction**

744 Likely doublets or multiplets were detected and carefully removed through a multi-step approach
745 as described in our recent studies[15,41]. Briefly, doublets or multiplets were identified by the
746 following methods: 1) Library complexity: cells with high-complexity libraries in which detected
747 transcripts are aligned to > 6,500 genes were removed. 2) Cluster distribution and marker gene
748 expression: some doublets or multiplets can form distinct clusters with hybrid expression
749 features, exhibiting an aberrantly high gene count. We reviewed the expression levels and
750 proportions of canonical lineage-related marker genes in each identified cluster. Clusters co-
751 expressing discrepant lineage markers were identified and removed. 3) Doublet detection
752 algorithms: we applied DoubletFinder[42], an algorithm to predict doublets in scRNA-seq data, to

753 detect doublets. The proportion of expected doublets was estimated based on the number of cells
754 used for scRNA-seq library construction. Library size normalization was then performed using
755 Seurat[43] on the filtered gene-cell matrix to obtain normalized UMI count. Statistical assessment
756 of possible batch effects was performed using the R package k-BET[44], as described in our
757 previous studies[15,41] and Harmony[18] was run with default parameters to remove batch effects
758 present in the PCA space.
759
760 **Unsupervised clustering and subclustering analysis**
761 Seurat[43] was applied to the normalized gene-cell matrix to identify highly variable genes (HVGs)
762 for unsupervised cell clustering. Principal component analysis (PCA) was performed on the top
763 2,000 HVGs. The elbow plot was generated with the *ElbowPlot* function of Seurat and, based on
764 which, the number of significant principal components (PCs) was determined. The
765 *FindNeighbors* function of Seurat was used to construct the Shared Nearest Neighbor (SNN)
766 Graph, based on unsupervised clustering performed with Seurat function *FindClusters*. Multiple
767 rounds of clustering and subclustering analysis were performed to identify major cell types and
768 distinct cell transcriptional states. Dimensionality reduction and 2-D visualization of the single
769 cell clusters was performed using Uniform Manifold Approximation and Projection (UMAP)[45]
770 with Seurat function *RunUMAP*. The number of PCs used to calculate the embedding was the
771 same as that used for clustering. To determine whether cells are over-clustered, we checked for
772 each cell type, whether multiple subclusters represent the same cell state without showing any
773 unique features. In addition, we applied ROGUE[38], an entropy-based statistic, to quantify the
774 cellular transcriptional heterogeneity of each cluster. Subclustering analysis was then performed
775 for low-purity clusters suggested by ROGUE. Hierarchical clustering of major epithelial subsets

776 was performed on the Harmony batch-corrected PCA dimension reduction space. The
777 hierarchical tree was computed based on the Euclidean distance with Ward linkage. Dendrogram
778 was generated using the R function *plot.hc*.
779
780 **Identification of malignant cells**
781 Malignant cells were distinguished from non-malignant subsets based on information integrated
782 from multiple sources as described in our recent studies[15,41]: 1) Cluster distribution: due to the
783 high degree of inter-patient tumor heterogeneity, malignant cells often exhibit distinct cluster
784 distribution from that of normal epithelial cells. While non-malignant cells derived from
785 different patients often are clustered together by cell type, malignant cells from each patient's
786 tumor likely form a separate cluster. 2) Copy-number alterations (CNAs): We applied
787 inferCNV[17] (version 1.3.2) to infer large-scale CNAs in each individual cell with T cells as the
788 reference control to separate malignant cells from non-malignant cells. To quantify the cell level
789 copy number variations, CNV score were aggregated using a strategy similar to that described in
790 a previous study[17]. Briefly, we first computed arm-level CNV scores as the mean of the squares
791 of CNV values across each chromosomal arm. The arm-level CNV scores were further
792 aggregated across all chromosomal arms by taking the arithmetic mean value of the arm-level
793 scores using R function *mean*. 3) Marker gene expression: we also examined the expression of
794 lung epithelial lineage-specific genes and LUAD-related oncogenes in identified epithelial cell
795 clusters. 4) Cell level expression of $KRAS^{G12D}$ mutations. To extract the single cell somatic
796 variant for *KRAS* mutation in patients, the alignment records were extracted from the original
797 BAM file using the mutation location information. Unique mapping alignments (MAPQ=255)
798 labelled as either PCR duplication or secondary mapping were filtered out. The resulting somatic

799 variant carrying reads were evaluated using IGV[46] and the corresponding cell-specific barcodes
800 were then extracted and used for downstream analysis.
801
802 **PCA analysis of malignant cells**
803 Raw unique molecular identifier counts of identified malignant cells were log normalized and
804 used for PCA analysis using Seurat (*RunPCA* function). The PCA dimension reduction data was
805 extracted using *Embeddings* function. The top three most highly ranked PCs were exported for
806 visualization using JMP v15. The 3D scatterplot of PCA data was generated using the scatterplot
807 3D tool in JMP v15.
808
809 **Determination of non-malignant cell types and states**
810 The non-malignant cell types and states were determined based on unsupervised clustering
811 analysis following batch effects correction using Harmony[18]. Two rounds of clustering analysis
812 were performed on non-malignant cells to identify major cell types and cell transcriptional states
813 within the major cell types. The clustering and UMAP visualization of normal epithelial cells
814 (**Fig. 1b**) was performed using Seurat with the default parameters. Specifically, we used k.param
815 = 20 and resolution = 0.4 for the shared nearest-neighbor (SNN) graph construction and cluster
816 identification, respectively. UMAP visualization was performed with the default parameters
817 (min.dist = 0.3). For clustering analysis of airway epithelial cells and alveolar epithelial cells,
818 RunPCA function was used to determine the most contributing top PCs in each sub-population.
819 Same clustering approaches were used k.param = 20 and resolution = 0.4 for SNN graph
820 construction and clusters identification. UMAP plots were generated with min.dist = 0.3 using
821 *RunUMAP* function in Seurat. Density plots of alveolar intermediate cells (**Fig. 3h**) was

822 generated using *stat_densit_2d* function in R package ggplot2 (version: 3.3.5) with the first two
823 UMAP dimension reduction data as the input. For normal epithelial cells (fig 1.b), differentially
824 expressed genes (DEGs) of each cluster were identified using the *FindMarkers* function in
825 Seurat with the following parameters: FDR adjusted $p < 0.05$ and fold change > 1.2. We made
826 references to the expression of canonical epithelial marker genes of previously published
827 studies[15,47,48] to define the normal epithelial cell types and states. Bubble plots were generated for
828 selected DEGs and canonical markers to define AT1 (*AGER1*/*ETV5/PDPN*+), AT2
829 (*SFTPB*/*SFTPC*/*ETV5*+), bronchioalveolar (*SFTPC*/*SFTPB*+ and *SCGB1A1*/*SCGB3A1*+), AIC
830 (*AGER1*/*ETV5/PDPN*+ *and SFTPB*/*SFTPC*+), club and secretory
831 (*SCGB1A1*/*SCGB3A1/CYP2F1*+), basal (*KRT5*/*TP63*+), ciliated (*CAPS*/*PIFO*/*FOXJ1*+),
832 ionocyte (*ASCL3*/*FOXI1*+), neuroendocrine (*CALCA*/*ASCL1*+) and tuft
833 (*ASCL2*/*MGST2*/*PTGS1*+). We identified KACs as a subset of AICs by unsupervised clustering
834 of AICs and defined them based on marker genes identified in previous reports[30,31,49] including
835 significant up regulation of the following genes relative to other alveolar cells: *KRT8*, *CLDN4*,
836 *PLAUR* as well as *CDKN1A*/*2A*.
837
838 **Scoring of curated gene signatures**
839 We measured the expression levels of previously defined ITH meta-programs (MPs) by Gavish
840 *et al*.[21] which analyzed 1456 samples covering 24 cancer types and profiling 2,591,545 cells.
841 Among a total of 41 consensus ITH MPs identified, MPs with unassigned functional annotations
842 (Unassigned MPs 38-41; n = 4), neural/ hematopoietic lineage specific MPs (MPs 25-29, MPs
843 33-37; n = 10) as well as cell type specific MPs irrelevant to LUAD (MPs 22-24 secreted/cilia,
844 MP 32 skin-pigmentation; n = 4) were filtered out, resulting in 23 MPs which intimately

845 correlated with the hallmarks of cancer[50] for further analysis. The signature scores were
846 computed using the *AddModuleScore* function in R package *Seurat*, similarly as is described
847 previously[15,41]. The *KRAS* signature used in this study was derived by calculating the DEGs
848 between *KRAS*-mutant malignant cell enriched cluster and other malignant cells (FDR adjusted
849 p-value < 0.05, log fold-change > 1.2). The KACs signature used in this study (human and
850 mouse) was derived by calculating the DEGs between KACs and other alveolar cells (FDR
851 adjusted p-value < 0.05, log fold-change > 1.2). The KACs signature score and *KRAS* signature
852 score were calculated using the *AddModuleScore* function in Seurat.
853
854 **Reconstruction of alveolar cell differentiation trajectories**
855 Reconstruction of the differentiation trajectory of alveolar cells and malignant cells were
856 performed using Monocle 2[51] by inferring the pseudotemporal ordering of cells according to
857 their transcriptome similarity. Monocle 2 analysis of malignant cells from P14 was performed
858 with default parameters on the robustly expressed genes identified using the *detectGenes*
859 function with default parameters. Detected genes were further required to be expressed by at
860 least 50 cells. To show the expression dynamics of ITH meta-programs[21] as pseudotime
861 progresses, ITH meta-program scores were plotted along the pseudotime axis and the smoothed
862 lines was generated using the smoother tool in JMP Pro v15. CytoTRACE[20] was applied with
863 default parameters to infer the cellular differentiation states to complement the trajectory analysis
864 and robustly understand the cellular differentiation hierarchy. The *normalmixEM* function from
865 mixtools R package was used to determine the threshold of the CytoTRACE score in AICs
866 (**Extend Data Fig. 3c**) with k = 2. The final threshold calculated for dichotomizing the AICs into
867 high and low differentiation groups is 0.58.

868

869 **Survival analysis**

870 Survival analysis including overall survival (OS) status and survival time from TCGA LUAD
871 cohort was performed as is previously described[15]. *KRAS* mutation status in TCGA LUAD
872 samples was downloaded from cBioPortal (https://www.cbioportal.org, study ID:
873 luad_tcga_pan_can_atlas_2018). We used the log-rank test to calculate *P*-values between groups,
874 and the Kaplan–Meier method to generate survival curves. The statistical significance testing in
875 this analysis was two-sided. To control for multiple hypothesis testing, we applied the
876 Benjamini–Hochberg method to correct *P*-values and the FDR Q values were calculated where
877 applicable. Results were considered statistically significant at *P*-value or FDR Q value of < 0.05.

878

879 **Bulk RNA extraction and WES**

880 Total RNA was isolated from homogenized cryosections of human lung tissues and, when
881 available, from frozen PBMCs using Qiagen RNeasy mini kit (Qiagen) according to
882 manufacturer's recommendations. Qubit 4 Fluorometer (Thermo Fisher Scientific, Waltham,
883 MA) was used for measurement of RNA yield, and RNA quality was assessed by NanoDrop
884 1000 (Thermo Fisher Scientific) as well as the Agilent RNA 6000 Pico Kit (Agilent, Santa Clara,
885 CA) for computation of RNA integrity numbers. Whole exome sequencing (WES) was
886 performed to analyze recurrent driver mutations, using PBMCs, or most distant NL tissues when
887 blood draw was not consented, as germline control. WES data were processed and mapped to
888 human reference genome, somatic mutations were identified, annotated as previously
889 described[52,53]with further filtration steps. Briefly, only MuTect[54] calls marked as "KEEP" were
890 selected and taken into the next step. Mutations with a low variant allelic fraction (VAF<0.02) or

891 had a low alt allele read coverage (< 4) were removed. After that, common variants reported by
892 ExAc (the Exome Aggregation Consortium, http://exac.broadinstitute.org), Phase-3 1000
893 Genome Project (http://phase3browser.1000genomes.org/Homo_sapiens/Info/Index), or the
894 NHLBI GO Exome Sequencing Project (ESP6500) (http://evs.gs.washington.edu/EVS/) with the
895 minor allele frequency greater than 0.5% were further removed. The intronic mutations,
896 mutations at 3' or 5' UTR or UTR flanking regions, silent mutations were also removed. The
897 mutation load was calculated as the number of nonsynonymous somatic mutations (nonsense,
898 missense, splicing, stop gain, and stop loss substitutions and frameshift insertions and deletions)
899 called in a sample.
900
901 **Statistical analyses**
902 In addition to the algorithms and statistical analyses described above, all other basic statistical
903 analysis was performed in the R statistical environment (version 4.0.0). All statistical tests
904 performed in this study were two-sided. The Kruskal–Wallis H test was used to compare
905 variables of interests across three or more groups. Wilcoxon Rank-Sum test was used for paired
906 comparisons among matched samples from the same patients. Wilcoxon Rank-Sum test was also
907 used to compare other continuous variables such as gene expression levels, signature scores
908 between groups. The Spearman's correlation coefficient was calculated to assess the association
909 between two continuous variables (e.g., the cellular proportions and gene signature scores).
910 Fisher's exact test was used to identify enriched categories between two categorical variables.
911 For the TCGA dataset, the clinical data were downloaded from the PanCanAtlas study[20].
912 Benjamin-Hochberg method was used to control for multiple hypothesis testing. All statistical
913 significance testing in this study was two-sided and results were considered statistically

914 significant at *P*-values or FDR *q*-values < 0.05. When a *P*-value reported by R was smaller than
915 2.2e-16, it was reported as "$P < 2.2 \times 10^{-16}$".
916
917 **Digital spatial imaging of human tissues**
918 We used standard immunofluorescence with the following antibodies: Claudin 4 (Clone 3E2C1,
919 AF594, LSBio, catalog number LS-C354893 concentration 0.5 µg/ml and Keratin 8 (Clone
920 EP1628Y, AF647, Abcam, catalog number ab192468, concentration 0.25 µg/ml). Optimization
921 of immunofluorescence biomarkers was previously performed with different dilutions using
922 colorectal carcinoma and lung adenocarcinoma tissue. Immunofluorescence staining was
923 performed with standard protocol for Morphology Biomarkers only using the GeoMx Digital
924 Spatial profiling protocol. Slides were scanned at 20X using the GeoMx Digital Spatial profiling
925 platform. Following completion of the scanning, multiplex immunofluorescence image slides
926 were visualized adjusting channel thresholds for each fluorophore. Biomarker expression was
927 assessed in the following histological areas: Adenocarcinoma cells, adjacent reactive lung tissue
928 and distant non-reactive lung tissue.
929
930 **Animal housing and tobacco carcinogen exposure experiments**
931 Animal experiments were conducted according to Institutional Animal Care and Use
932 Committee–approved protocols. Mice were maintained in a pathogen-free animal
933 facility. *Gprc5a*$^{-/-}$ mice were generated as previously described[24,55]. Sex- and age-
934 matched *Gprc5a*$^{-/-}$ mice were divided into starting groups of 12 mice per exposure (NNK or
935 saline control) and timepoint (end-of-exposure or 7 months post-exposure, n = 48 mice). 8-week
936 old mice were intraperitoneally injected with 50 mg/kg of body weight nicotine-specific

937 nitrosamine ketone (NNK, Toronto Research Chemicals) or vehicle 0.9% saline (control), three
938 times per week for 8 weeks. At the end-of-exposure or at 7 months post-exposure, lungs were
939 harvested for derivation of live single cells for scRNA-seq (4 mice/group) or were collected in
940 OCT for immunofluorescence (IF) analysis.
941
942 *Sftpc*$^{CreER/+}$; *Rosa*$^{Sun1GFP/+}$ were generously provided by Dr. Harold Chapman (University of
943 California, San Francisco) and were crossed to *Gprc5a*$^{-/-}$ mice to generate *Sftpc*$^{CreER/+}$;
944 *Rosa*$^{Sun1GFP/+}$; *Gprc5a*$^{-/-}$ mice for AT2-derived organoid experiments. *Sftpc*$^{CreER/+}$; *Rosa*$^{Sun1GFP/+}$;
945 *Gprc5a*$^{-/-}$ mice were treated with 50 mg/kg NNK or control saline (IP) (n = 6 each), three times
946 per week for 8 weeks. At week 6 of treatment (two weeks before the end of exposure), mice from
947 both groups received 250 µg (IP) Tamoxifen dissolved in corn oil for four consecutive days.
948 Lungs were harvested at the end of saline or NNK exposure and digested to derive live single
949 cells for organoid experiments.
950
951 **Chemicals and reagents**
952 Tobacco-specific carcinogen (nicotine-specific nitrosamine ketone; NNK) with a purity of
953 99.96% by HPLC was purchased from Toronto Research Chemicals (North York, ON, Canada).
954 Tamoxifen, as well as Hematoxylin and Eosin (H&E) staining reagents, were purchased from
955 Sigma Aldrich (St. Louis, MO).
956
957 **Single-cell isolation from mouse lung tissues**
958 Mouse lungs were collected at end-of-exposure or 7 months post exposure to NNK or saline and
959 freshly digested for derivation of viable single-cells. Mouse lungs were immediately collected in

960 6 cm dish containing ice-cold Leibovitz on ice. Extra-pulmonary airways were removed, and
961 lungs were transferred to a new dish on ice and minced into <1mm pieces with curved scissors.
962 Using wide-orifice low retention tips, tissue pieces were collected in freshly prepared ice-cold
963 enzyme solution (1.35 ml 1x Leibovitz + 150 µl of 20 mg/ml Collagenase Type I + 150 µl of 20
964 mg/ml Elastase + 40 µl 20 mg/ml DNaseI) and transferred to 5 ml LoBind Eppendorf tube.
965 Tissues were digested and RBCs were removed as described above for human tissues. Murine
966 single-cell were immediately stained with SYTOX blue, CD45-PE/Cy7 (BioLegend 103114),
967 Icam2-A647 (Life Technologies A15452), Epcam-BV421 (BioLegend 118225), and Ecad-A488
968 (eBioscience 53-3249-80). Viable (SYTOX Blue negative) epithelial (CD45-, Icam2-, Epcam+
969 and/or Ecad+) cells were sorted using FACS Aria flow cytometer and collected into PBS + 2%
970 FBS on ice. Cells were washed again to eliminate ambient RNA, and a sample was taken for
971 counting by Trypan Blue exclusion (T8154, Sigma Aldrich). Up to 10,000 cells per sample were
972 loaded on 10X Chromium microfluidic chips, followed by generation of single-cell libraries in
973 the manner described above for cells derived from human samples.
974
975 **Histopathological and IF analysis of mouse lung tissues**
976 After sacrifice, lungs were inflated by injection with 50% optimal cutting temperature (OCT)
977 reagent in PBS, excised for histopathological assessment. Lung surface lesions were examined
978 by macroscopic observation and immediately cryopreserved on dry ice. H&E stained
979 cryosections were digitally scanned using the Aperio ScanScope Turbo slide scanner (Leica
980 Microsystems Inc., Buffalo Grove, IL) at 200X magnification, and visualized by ImageScope
981 software (Leica Microsystems, Inc.). Lung sections from all mice were analyzed for lung tumor
982 incidence according to previously reported criteria that include assessment of tumor

983 differentiation and invasiveness as well as nuclear crowding, size, and atypia[56]. Total lung area,
984 as well as area of each of the tumor mass and malignant tumor lesions were selected and
985 measured by two experienced pathologists (JF) using the Aperio Image Toolbox (Leica
986 Microsystems Inc.). Lung tumors were enumerated from each group of mice by histopathological
987 analysis of H&E stained slides and according to previously reported criteria for characterization
988 of murine lung lesions[56]. Tumor volume (mm^3) was calculated by multiplying the malignant
989 tumor area (mm^2) by the thickness of tissue sections on each slide. Tumor burden was then
990 defined as percentage of tumor volume from the entire lung volume. Cryofixed lung tissue
991 sections were obtained for IF analysis of Lamp3 (clone 391005, Synaptic Systems), Krt8
992 (TROMA-I clone from the University of Iowa DSHB), and Pdpn (clone 8.1.1, from the
993 University of Iowa DSHB). Images were acquired using an immunofluorescence microscope.
994

995 **3D culture and analysis of AT2-derived organoids**

996 *Sftpc*$^{CreER/+}$; *Rosa*$^{Sun1GFP/+}$; *Gprc5a*$^{-/-}$ treated with NNK or saline were sacrificed at end-of-
997 exposure. Lungs were harvested, dissociated into single cells, and live GFP+ single cells were
998 collected by flow cytometry using a FACS Aria II instrument as previously described[57]. GFP^+
999 AT2 cells from NNK- or saline-treated groups were immediately washed and re-suspended at a
1000 concentration of 5,000 cells/50 µl of 3D media (F12 medium supplemented with
1001 insulin/transferrin/selenium, 10% FBS, penicillin/streptomycin, L-glutamine). GFP+ cells were
1002 mixed at a 1:1 ratio with 50,000 mouse endothelial cells (harvested from mouse lungs by CD31
1003 selection and expanded *in vitro* as previously described [58]), and resuspended in 50 µl of
1004 Geltrex™ reduced growth factor basement membrane matrix (Gibco, A1413301). 100 µl of 1:1
1005 GFP^+:endothelial cell mixture was plated on transwell inserts with 0.4 µm pores and allowed to

1006 solidify for 30 mins in a humidified CO_2 incubator. Each well was then supplemented with 3D
1007 media containing ROCK inhibitor (Millipore, Y-27632) and recombinant mouse FGF-10 (R&D
1008 Systems, 6224-FG), and plates were incubated at 37°C in a humidified CO_2 incubator. Wells
1009 were replenished with 3D media every other day. At day 21 post-plating, 3D organoids were
1010 harvested from the basement membrane matrix using Gentle Cell Dissociation Reagent
1011 (StemCell Technologies, 100-0485), fixed with 4% paraformaldehyde, permeabilized, blocked,
1012 and stained overnight at 4°C with a mixture of primary antibodies for immunofluorescence
1013 analysis: guinea pig anti-lysosomal associated membrane protein 3 (LAMP3, 1:1000, Synaptic
1014 Systems, 391005), chicken anti-green fluorescent protein (GFP, 1:5000, Abcam, AB13970), rat
1015 anti-keratin8 (Krt8, 1:100, Developmental Studies Hybridoma Bank, TROMA-I). The next day,
1016 organoids were washed and stained with fluorophore-conjugated secondary antibodies overnight
1017 at 4°C while being protected from light. DAPI nuclear stain was added to the secondary antibody
1018 mix for 30 minutes, after which organoids were washed, collected in Aqua-Poly/Mount
1019 (Polysciences, 18606-20) and transferred to slides. Images of organoids were captured using an
1020 Andor Revolution XDi WD Spinning Disk Confocal microscope and analyzed using Imaris
1021 software.
1022
1023 **Data Availability**
1024 Sequencing data for P1-P5 were previously generated and deposited in the European Genome–
1025 phenome Archive (EGA) under the accession number EGAS00001005021[15]. All other
1026 sequencing data generated in this study (human and mouse) will be deposited in EGA and
1027 publicly available upon manuscript acceptance.
1028

1029

1030 **Code Availability**

1031 Detailed codes for all analyses will be shared via GitHub (https://github.com/) and publicly

1032 available very soon.

1033

1034

1035 **Methods references**

1036

1037 40 Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen

1038 human tumors. *Nature Medicine* **26**, 792-802, doi:10.1038/s41591-020-0844-1 (2020).

1039 41 Wang, R. *et al.* Single-cell dissection of intratumoral heterogeneity and lineage diversity

1040 in metastatic gastric adenocarcinoma. *Nature Medicine* **27**, 141-151, doi:10.1038/s41591-

1041 020-1125-8 (2021).

1042 42 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in

1043 Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-

1044 337.e324, doi:10.1016/j.cels.2019.03.003 (2019).

1045 43 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell

1046 transcriptomic data across different conditions, technologies, and species. *Nature*

1047 *biotechnology* **36**, 411-420, doi:10.1038/nbt.4096 (2018).

1048 44 Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for

1049 assessing single-cell RNA-seq batch correction. *Nature methods* **16**, 43-49,

1050 doi:10.1038/s41592-018-0254-1 (2019).

1051 45 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and
1052 projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
1053 46 Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant
1054 Review with the Integrative Genomics Viewer. *Cancer Res* **77**, e31-e34,
1055 doi:10.1158/0008-5472.Can-17-0337 (2017).
1056 47 Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse
1057 brain. *Science* **367**, doi:10.1126/science.aay5947 (2020).
1058 48 Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA
1059 sequencing. *Nature* **587**, 619-625, doi:10.1038/s41586-020-2922-4 (2020).
1060 49 Maroni, G. *et al.* Identification of a targetable KRAS-mutant epithelial population in non-
1061 small cell lung cancer. *Commun Biol* **4**, 370, doi:10.1038/s42003-021-01897-6 (2021).
1062 50 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-
1063 674, doi:10.1016/j.cell.2011.02.013 (2011).
1064 51 Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis.
1065 *Nature* **566**, 496-502, doi:10.1038/s41586-019-0969-x (2019).
1066 52 Wang, R. *et al.* Multiplex profiling of peritoneal metastases from gastric adenocarcinoma
1067 identified novel targets and molecular subtypes that predict treatment response. *Gut* **69**,
1068 18-31, doi:10.1136/gutjnl-2018-318070 (2020).
1069 53 Hao, D. *et al.* Integrated genomic profiling and modelling for risk stratification in patients
1070 with advanced oesophagogastric adenocarcinoma. *Gut* **70**, 2055-2065,
1071 doi:10.1136/gutjnl-2020-322707 (2021).

1072 54 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and
1073 heterogeneous cancer samples. *Nature biotechnology* **31**, 213-219, doi:10.1038/nbt.2514
1074 (2013).
1075 55 Tao, Q. *et al.* Identification of the retinoic acid-inducible Gprc5a as a new lung tumor
1076 suppressor gene. *J Natl Cancer Inst* **99**, 1668-1682, doi:10.1093/jnci/djm208 (2007).
1077 56 Nikitin, A. Y. *et al.* Classification of proliferative pulmonary lesions of the mouse:
1078 recommendations of the mouse models of human cancers consortium. *Cancer Res* **64**,
1079 2307-2316 (2004).
1080 57 Barkauskas, C. E. *et al.* Type 2 alveolar cells are stem cells in adult lung. *J Clin Invest*
1081 **123**, 3025-3036, doi:10.1172/jci68782 (2013).
1082 58 Lee, J. H. *et al.* Lung stem cell differentiation in mice directed by endothelial cells via a
1083 BMP4-NFATc1-thrombospondin-1 axis. *Cell* **156**, 440-455,
1084 doi:10.1016/j.cell.2013.12.039 (2014).

1085

1086

1087 **Acknowledgements**

1088 We appreciate all the patients who generously provided their samples for this research. We thank
1089 Dr. Harold Chapman for kindly providing *Sftpc*$^{CreER/+}$; *Rosa*$^{Sun1GFP/+}$ mice. This work was
1090 supported in part by funding from Johnson and Johnson (to H.K.), the National Cancer Institute
1091 (NCI) grants R01CA205608 (to HK), 1U2CCA233238 (to H.K., S.M.D., and A.E.S.), University
1092 of Texas SPORE in Lung Cancer P50CA070907 and the P50 core grant CA016672, as well as
1093 the Cancer Prevention and Research Institute of Texas grants RP150079 (to H.K.), RP220101 (to

H.K.), and RP160668 (to I.I.W.). The sequencing core facility at MD Anderson Cancer Center is supported by the National Institutes of Health grant 1S10OD024977.

Author information

G.H., A.S., W.T., W.L., A.G.S, S.H.D., Z.R., X.C, E.D., L.I.B, E.R.P., S.J.M., J.C., J.F., P.S., I.I.W., L.W., and H.K. performed data curation and analysis. G.H., A.S., W.L., A.G.S, S.H.D., E.D., J.F., L.S., L.W., and H.K. developed methodology. G.H., A.S., W.L., A.G.S, S.D.H., L.S., L.W., and H.K. performed data validation. G.H., A.S., X.C, E.D., S.J.M.,J.C., S.M.D., P.S., I.I.W., C.S.S., A.E.S., L.W., and H.K. performed data visualization and interpretation. M.V.N., T.C., B.S., I.I.W., C.S.S., A.E.S., L.W., and H.K. provided resources for the work. C.S.S., A.E.S., L.W., and H.K. were involved in administration of the project. G.H., A.S., L.W. and H.K. wrote the original draft. C.S.S., A.E.S., L.W., and H.K. supervised the study. L.W. and H.K. conceived the study. All authors reviewed and/or edited the final version.

Ethics declarations

Competing interests

CCS and AES are employees of Johnson and Johnson. HK reports research funding from Johnson and Johnson. MVN receives research funding to institution from Mirati, Novartis, Checkmate, Alaunos/Ziopharm, AstraZeneca, Pfizer, Genentech, and consultant/advisory board fees from: Mirati, Merck/MSD, Genentech. TC reports speaker fees/honoraria from The Society for Immunotherapy of Cancer, Bristol Myers Squibb, Roche, Medscape, and PeerView; advisory role/consulting fees from AstraZeneca, Bristol Myers Squibb, EMD Serono, Merck & Co.,

Genentech, and Arrowhead Pharmaceuticals; and institutional research funding from AstraZeneca, Bristol Myers Squibb, and EMD Serono. SJM reports funding from Arrowhead Pharma and Boehringer Ingelheim outside the scopes of submitted work. BS reports speaker fees from AstraZeneca and PeerView, advisory fees from AstraZeneca, and consultation fees from Medscape. IIW reports grants and personal fees from Genentech/Roche, grants and personal fees from Bayer, grants and personal fees from Bristol-Myers Squibb, grants and personal fees from AstraZeneca, grants and personal fees from Pfizer, grants and personal fees from HTG Molecular, personal fees from Asuragen, grants and personal fees from Merck, grants and personal fees from GlaxoSmithKline, grants and personal fees from Guardant Health, personal fees from Flame, grants and personal fees from Novartis, grants and personal fees from Sanofi, personal fees from Daiichi Sankyo, grants and personal fees from Amgen, personal fees from Oncocyte, personal fees from MSD, personal fees from Platform Health, grants from Adaptive, grants from Adaptimmune, grants from EMD Serono, grants from Takeda, grants from Karus, grants from Johnson and Johnson, grants from 4D, from Iovance, from Akoya, outside the submitted work.

Additional information

Supplementary Information is available for this paper.

Extended data figure

1141 **Extended Data Fig. 1. Summary metrics for data quality, identification of malignant cells**
1142 **and screening of driver mutations in LUAD samples. a,** Number of cells retained for scRNA-
1143 seq in each sample. Y-axis is log-scaled. **b,** Fraction of mitochondrial genes (left) and the
1144 number of detected genes per cell (right). Box, median ± interquartile range; whiskers, 1.5×
1145 interquartile range; center line: median. **c,** Schematic plot illustrating the approach used to
1146 identify malignant cells in LUAD samples based on clustering, CNV, and expressed $KRAS^{G12D}$
1147 mutations (Expr mutation). **d,** Recurrent driver mutations identified by whole exome sequencing.
1148 **e,** UMAP plot of 17,064 malignant cells colored by smoking status.
1149
1150 **Extended Data Fig. 2. Characterization of inter- and intra-tumor heterogeneity of LUAD**
1151 **malignant cells. a,** Harmony-corrected UMAP plot of malignant cells colored by cluster ID
1152 (left) and cluster distribution by sample (right). **b**, Enrichment of hierarchical clusters from Fig.
1153 2d by patient. **c,** expression of 4 representative MP signatures among the 5 hierarchical clusters
1154 defined in Fig. 2d. Box, median ± interquartile range; whiskers, 1.5× interquartile range; center
1155 line: median. **d,** UMAP plots showing P14 malignant cells colored by expression of 3 MP
1156 signatures. **e**, Unsupervised clustering analysis of P14 malignant cells based on inferred CNV
1157 profiles (left). UMAP of P14 malignant cells (middle) and inferred trajectory (top right) colored
1158 by CNV clusters, as well as $KRAS^{G12D}$ mutation expression status along the developmental
1159 trajectory (bottom right). **f**, Signature score of alveolar lineage meta-program across the CNV
1160 clusters shown in panel e. ***: $p < 0.001$. *P*-value was calculated using Mann Whitney U test. **g**,
1161 Harmony-corrected UMAP plot of malignant cells colored by *KRAS* signature score (left).
1162 Correlation between MP30 and *KRAS* signature scores in malignant cells of KM-LUADs (right).
1163 *P*-values were calculated with Spearman correlation test. R denotes the Spearman correlation

1164 coefficient. **h**, Validation of *KRAS* signature score in TCGA LUAD cohort. Left: Kaplan-Meier
1165 plot showing the survival probability differences between samples with high level of *KRAS*
1166 signature, those with low level of *KRAS* signature, and those with $KRAS^{G12D}$ mutation. *KRAS* sig.
1167 high: samples within top quartile of *KRAS* signature score. *KRAS* sig. low: samples below the
1168 third quartile of *KRAS* signature score. mo: months. Right: heatmap showing the MP signature
1169 score of TCGA LUAD samples. *: $p < 0.05$. *P*-value was calculated with Chi-square test.
1170
1171 **Extended Data Fig. 3. Phenotypic diversity and states of normal epithelial cells. a,**
1172 Composition of normal epithelial lineages among the spatial regions at sample level. BA:
1173 bronchioalveolar. NE: neuroendocrine. Prolif: proliferative. **b**, Fractional changes of AT1, basal,
1174 ciliated, and club and secretory cells along the continuum of the spatial samples. Box, median ±
1175 interquartile range; whiskers, 1.5× interquartile range; center line: median. **c**, Distribution and
1176 composition of AICs with low (left) or high (right) CytoTRACE score. **d**, Differentially
1177 expressed genes between KACs and other AICs. **e,** Microphotographs of P10 (e) and P15 (f)
1178 LUADs (1-5) and paired uninvolved NL tissue (6-10). Top panels: Hematoxylin & Eosin
1179 staining showing LUAD tumor area (T) and tumor adjacent normal tissue (TAN; 1, top left),
1180 while 6 shows normal-appearing lung parenchyma. Immunofluorescence staining with KRT8 (2
1181 and 7, red), CLDN4 (3 and 8, yellow), PanCK (4 and 9, green), and composite image of all three
1182 antibodies (5 and 10). Blue nuclear staining was done using Syto13. Magnification 20x. Bar 200
1183 μm. **f,** AT1 signature expression across AT1, AT2, KACs and other AICs. Box definitions are
1184 the same as in panel **b**. **g,** Expression of Alveolar lineage signature in TCGA LUAD samples and
1185 matched NL tissues. ***: $p < 0.001$. *P*-value was calculated using Mann-Whitney U test. Box
1186 definitions are the same as in panel **b**. **h,** *KRT8* signature expression in malignant cells of KM-

1187 *versus KRAS*-wildtype LUADs. ***: *p*-value < 0.001. *P*-value was calculated using Mann-
1188 Whitney U test. Box definitions are the same as in panel **b**.
1189
1190 **Extended Data Fig. 4. Identification of KACs in tobacco carcinogenesis mouse model of**
1191 **KM-LUAD and the involvement of KACs in transition of AT2 to malignant cells during**
1192 **AT2-to-AT1 differentiation. a-b,** Expression levels of *Cd24a* and *Gkn2* in alveolar cells.
1193 Enlarged UMAPs and violin plots of alveolar cells colored by the expression level of *Cd24a* (a)
1194 and *Gkn2* (b). Box, median ± interquartile range; whiskers, 1.5× interquartile range; center line:
1195 median. **c**, Proportions of normal epithelial cell lineages and malignant cells in each sample. **d**,
1196 Fractional changes of malignant cells, AICs, AT2 and AT1 cells between EOE and 7 months
1197 post NNK treatment. whiskers, 1.5× interquartile range; center dot: median. **e**,
1198 Immunofluorescence analysis of Krt8 and Lamp3 in GFP-labelled AT2-derived mouse lung
1199 organoids following end of exposure to saline or NNK. Scale bar: 10 μm. **f**, KAC signature
1200 expression in samples of TCGA LUAD cohort (left) or preneoplasia cohort (right). AAH:
1201 atypical adenomatous hyperplasia. *P*-values were calculated using Mann-Whitney U test. Box,
1202 median ± interquartile range; whiskers, 1.5× interquartile range; center line: median.
1203
1204
1205 **Supplementary data table legends**
1206
1207 **Supplementary Data Table 1.** Sequencing data quality of all samples included in this study.
1208 Details of patient samples including sex, smoking status and sample location, as well as QC
1209 metrics of single-cell RNA-sequencing libraries in this study.

1210

1211 **Supplementary Data Table 2.** Fraction of epithelial cell lineages in each sample. Fraction of
1212 alveolar and airway lineages among total normal epithelial cells and across tumor and normal
1213 (tumor-adjacent, -intermediate, and -distant) samples of each patient.

1214

1215 **Supplementary Data Table 3.** Metaprogram scores across malignant cells (n = 17064). Scores
1216 of the 41 MPs from Gavish et al, CNV as well as differentiation, across individual malignant
1217 cells. Table also shows smoking and driver mutation status for each patient.

1218

1219 **Supplementary Data Table 4.** Significantly upregulated genes in KACs. List of top 50 most
1220 upregulated genes in human KACs. Average log fold change and adjusted *p* – values are
1221 indicated for each gene.

Figure 1

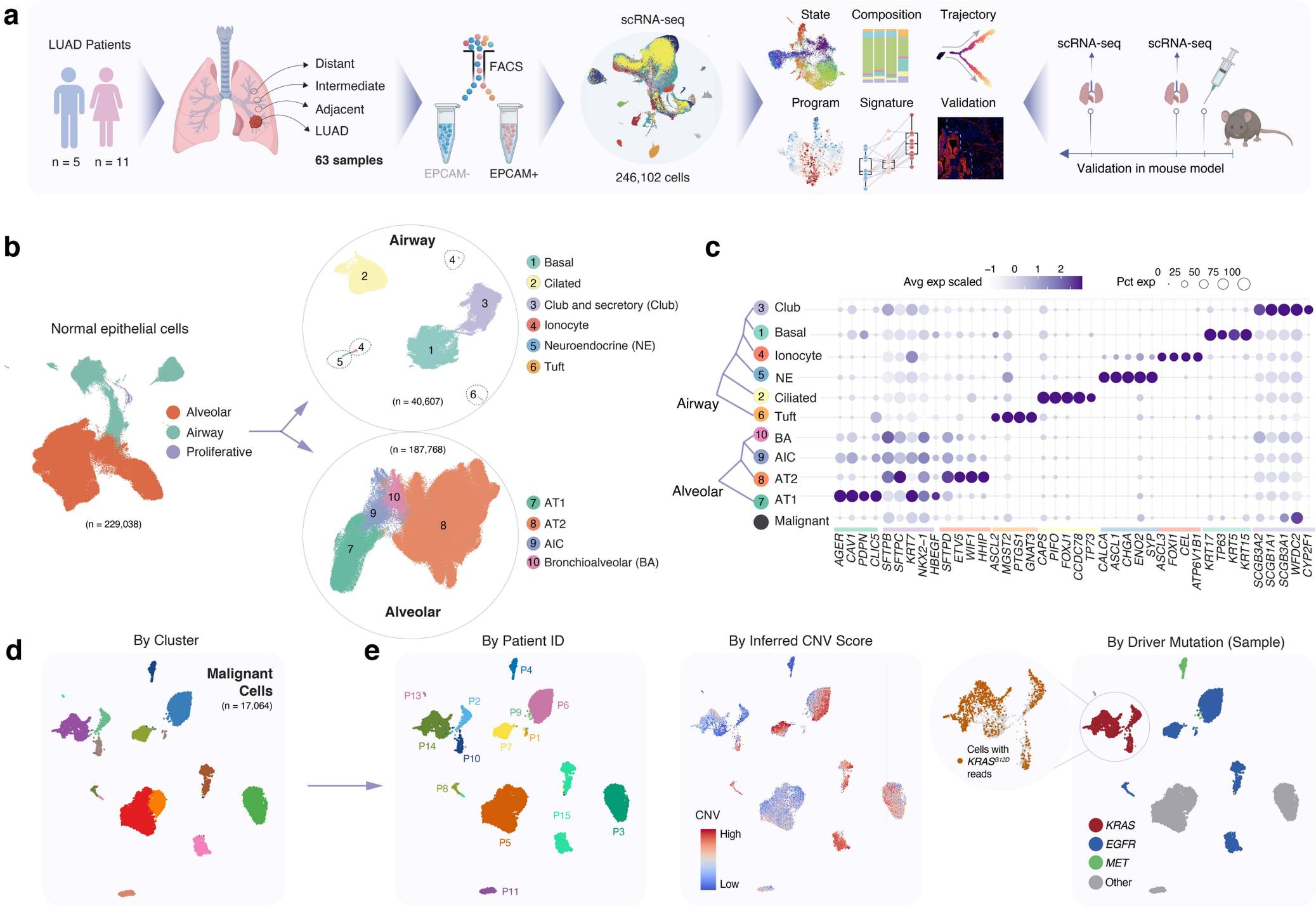

Fig. 1 | Transcriptional landscape of lung epithelial cells in early-stage LUAD patients. a, schematic overview of the experimental design and the analysis workflow. State: cellular transcriptional state. Composition: composition of cell subsets. Program: transcriptional programs in malignant cells. **b**, UMAP plots of 229,038 normal epithelial cells from 63 samples. Each dot represents a single cell colored by major cell lineages (left), airway sub lineages (top) and alveolar sub lineages (bottom). AIC: alveolar intermediate cell. **c**, Sub-lineage hierarchy within airways cells and alveolar cells (left) and expression (proportions and average expression levels) of selected marker genes for 10 normal epithelial cell types as defined in panel b (right) and malignant cells. **d**, Unsupervised clustering of 17,064 malignant cells colored by cluster ID. The zoom-in view (top left) showing malignant cells colored by *KRAS*^{G12D} mutation status identified using scRNA-seq data. **e**, UMAP plots of malignant cells as shown in **d**, but colored by patient ID (left), inferred CNV score (middle), and driver mutations identified in each tumor sample using whole-exome sequencing (right).

Figure 2

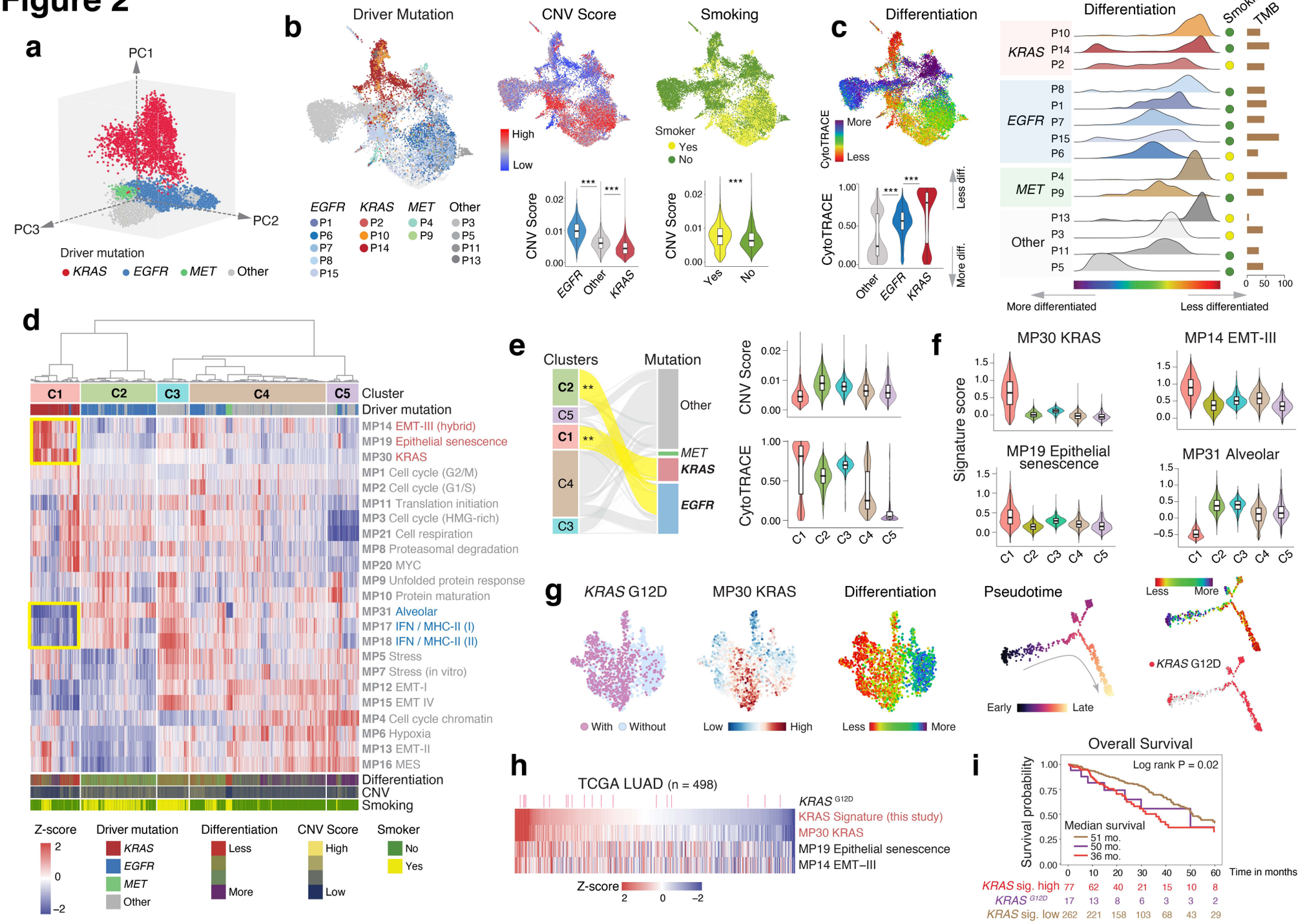

Fig. 2. Deep single-cell analysis of malignant cells reveals unique signatures, states, and malignant programs. a, Principal component analysis (PCA) plot of malignant cells colored by driver mutations identified in each tumor sample using whole-exome sequencing. **b**, Harmony-corrected UMAP plots of malignant cells colored by patient ID (left, which were assigned to color palette groups based on driver mutation status), CNV scores (top middle), smoking status (top right). Comparison of CNV scores between malignant cells from samples carrying different driver mutations (bottom left) or between smokers and never smokers (bottom right). ***: FDR adjusted p-value < 0.001. Boxes: 25th to 75th percentile, whiskers: 1.5× interquartile range, center line: median. **c**, Landscape of malignant cell differentiation status (top left). Comparison of differentiation states (inferred by CytoTRACE) between malignant cells from samples with different driver mutations (bottom left). Per sample distribution of cell CytoTRACE score right). TMB: tumor mutation burden. Boxes definitions are the same as in panel **b**. **d**, Unsupervised clustering of malignant cells based on expression of 23 previously defined consensus intratumor heterogeneity meta-programs (MPs). **e**, Enrichment of clusters (C1-C5) from panel **d** in samples with different driver mutations. Distribution of CNV scores (top right) and CytoTRACE scores (bottom right) across the 5 clusters defined in panel **d**. Boxes definitions are the same as in panel b. **f**, Distribution of signature scores of 4 representative MPS: across clusters defined in panel **d**. Boxes definitions are the same as in panel **b**. **g**, Profiling of intratumoral heterogeneity in malignant cells from P14. UMAP plots show P14 malignant cells colored by (left to right) *KRAS*^{G12D} mutation status, *KRAS* signature expression, and cell differentiation status (CytoTRACE), respectively. Developmental trajectory of P14 malignant cells colored by inferred pseudotime, differentiation status, and the presence of *KRAS*^{G12D} mutations. **h**, *KRAS* signature expressionin TCGA LUAD cohort. Enrichment of significantly correlated meta-programs with *KRAS* signature in TCGA LUAD cohort are shown. Samples harboring somatic *KRAS*^{G12D} mutations are annotated on the top. **i**, Survival probability differences between samples carrying somatic *KRAS*^{G12D} mutation, and those with high or low level of *KRAS* signature expression. *KRAS* sig. high: samples showing high *KRAS* signature expression (the top quartile) . *KRAS* sig. low: samples with low *KRAS* signature expression (below the third quartile). *KRAS*^{G12D}: samples with *KRAS*^{G12D} mutation.

Figure 3

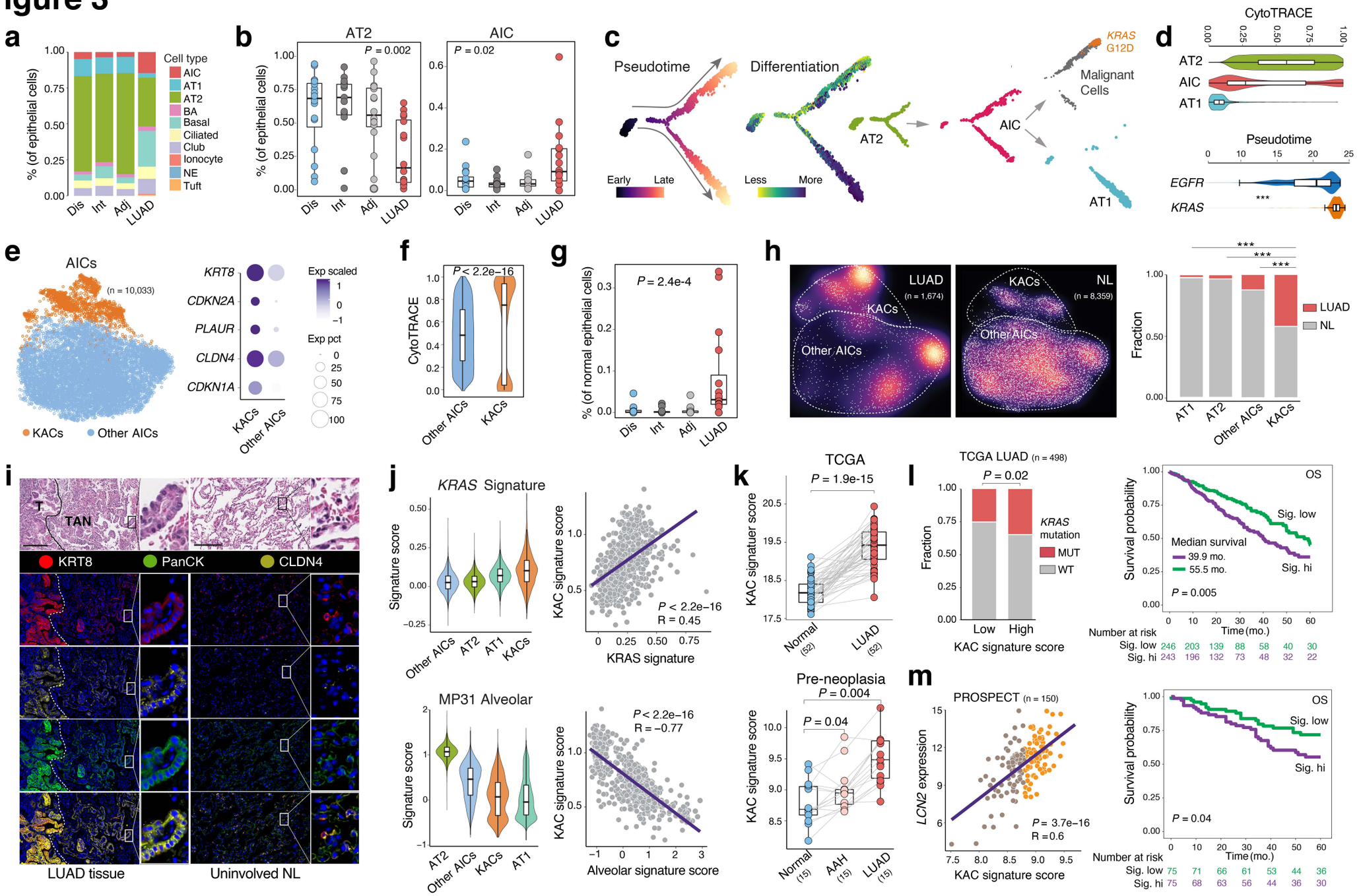

Fig. 3. I Identification and characterization of *KRT8*+ alveolar intermediate cells that are implicated in human LUAD development. a, Composition of normal epithelial lineages across spatial regions as defined in **Fig. 1a**. Dis: distant normal. Int: intermediate normal. Adj: adjacent normal. AIC: alveolar intermediate cells. AT1: alveolar type 1. AT2: alveolar type 2. BA: bronchioalveolar. NE: neuroendocrine. **b**, Changes in cellular fractional of AT2 cells (left) and AICs (right) across the spatial samples. *P*-values were calculated with Kruskal-Wallis test. median ± interquartile range; whiskers, 1.5× interquartile range; center line: median. **c**, Pseudotime trajectory analysis of alveolar and malignant cells identified in this study. **d**, Distribution of CytoTRACE scores in AICs, AT1 and AT2 cells (top). Distribution of pseudotime scores in malignant cells of *EGFR*- and *KRAS*-mutant tumors. ***: *p*-value < 0.001. P-value was calculated with Mann-Whitney U test. Boxes definitions are the same as in panel **b**. **e**, Sub-clustering analysis of AICs (left). KACs: *KRT8*+ alveolar intermediate cells. Expression levels and frequencies of representative marker genes of KACs (right). **f**, CytoTRACE score in KACs versus other AICs. Boxes definitions are the same as in panel **b**. **g**, Proportion of KACs among non-malignant epithelial cells across the spatial samples. P-value was calculated using Kruskal-Wallis test. Boxes definitions are the same as in panel **b**. **h**, Differences in cell density between LUAD (left) and normal lung tissues (middle). NL: normal lung. Barplot (right) shows fraction of alveolar cell subsets colored by sample type. ***: FDR adjusted *p*-value < 0.001. *P*-values were calculated using Fisher's exact tests. **i**, Microphotographs of P14 LUAD (1-5) and paired uninvolved NL tissue (6-10). Top panels: Hematoxylin & Eosin staining showing LUAD tumor area (T) and tumor adjacent normal tissue (TAN) displaying reactive hyperplasia of AT2 cells and presence of intra-alveolar macrophages (1, top left), while 6 shows TAN. Immunofluorescence staining with KRT8 (2 and 7, red), CLDN4 (3 and 8, yellow), PanCK (4 and 9, green), and composite image of all three antibodies (5 and 10). Blue nuclear staining was done using Syto13. Magnification 20x. Bar 200 μm. **j**, Enrichment of KRAS signature score (top left) and the alveolar lineage signature score (bottom left) across different subsets of alveolar cells. Correlation analysis between KAC signature scores and *KRAS* signature scores (top right) or alveolar lineage signature scores (bottom right). *P*-values were calculated with Spearman correlation test. R denotes the Spearman correlation coefficients. Boxes definitions are the same as in panel **b. k**, Enrichment of KAC signature score across samples of TCGA LUAD (top) and premalignancy (bottom) cohorts. AAH: atypical adenomatous hyperplasia. *P*-values were calculated using Wilcoxon signed-rank test.Boxes definitions are the same as in panel **b**. **l**, Fraction of *KRAS*-mutant samples among TCGA samples with high or low KAC signature score (left). MUT: *KRAS* mutant. WT: *KRAS* wild type. Low: LUAD samples with signature scores lower than the group median value. High: LUAD samples with signature scores higher than the group median value. Differences in overall survival probability between samples with high versus low KAC signature scores (right). Mo: months. **m**, Correlation between the expression levels of *LCN2* and KAC signature scores in the PROSPECT cohort (left). Overall survival probability differences between samples with high versus low KAC signature scores in PROSPECT cohort (right). Group median value of KAC signature scores was used dichotomize the high and low groups.

Figure 4

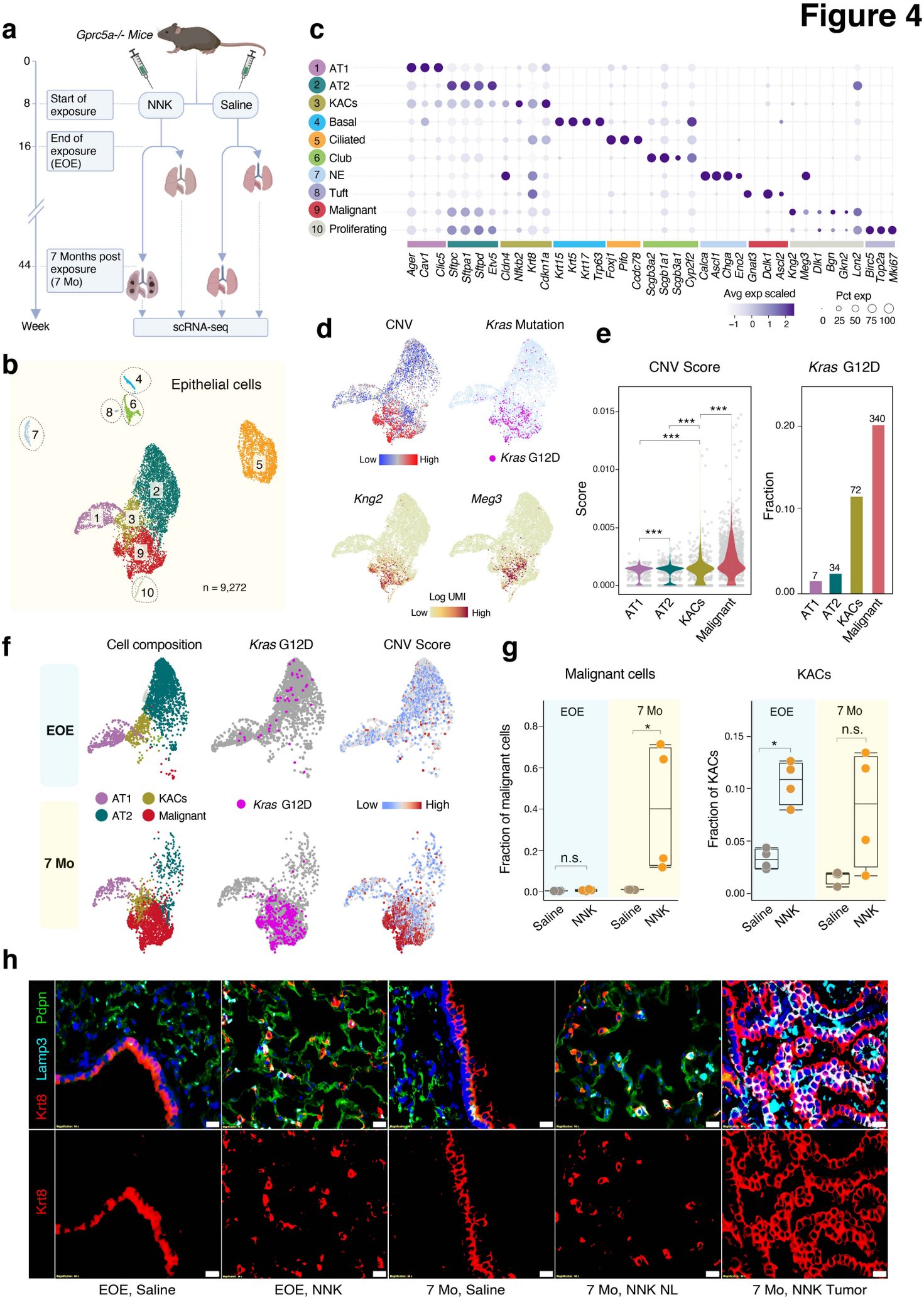

Fig. 4 | *Krt8+* alveolar intermediate cells evolve early and prior to tumor onset in vivo during tobacco-associated KM-LUAD pathogenesis. **a**, Schematic view of the experimental design. NNK: nicotine-derived nitrosamine ketone. **b**, UMAP view of mouse epithelial cell subsets. **c**, Proportions and average expression levels of selected marker genes for mouse normal epithelial cell lineages and malignant cells as defined in panel **a**. NE: neuroendocrine. KACs: *Krt8+* alveolar cells. **d**, UMAP distribution of alveolar and malignant cells colored by CNV score (top left), *Kras*^{G12D} mutation status (top right), and expression level of oncogenic markers *Kng2* and *Meg3*. **e**, Distribution of CNV scores (left) among alveolar and malignant cells and fraction of *Kras*^{G12D} mutant cells in KACs, malignant, AT1 and AT2 subsets (right). Absolute numbers of *Kras*^{G12D} mutant cells are indicated on top of each bar. **f**, Enlarged UMAPs of alveolar and malignant cells colored by cell subsets (left), *Kras*^{G12D} mutation status (middle) and CNV level (right), both at end-of-exposure to NNK (top row) and at 7 months post-NNK exposure (bottom row). EOE: end of exposure. Mo: months. **g**, Fraction of malignant cells (left) and KACs (right) across treatment groups and time points. *: FDR adjusted *p*-value < 0.05. P-values were calculated using Mann-Whitney U test. Box, median ± interquartile range; whiskers, 1.5× interquartile range; center line: median. h, Immunofluorescence of Krt8, Lamp3 and Pdpn level in mouse lung tissues. NL: normal lung. Scale bars: 10 µm.

Figure 5

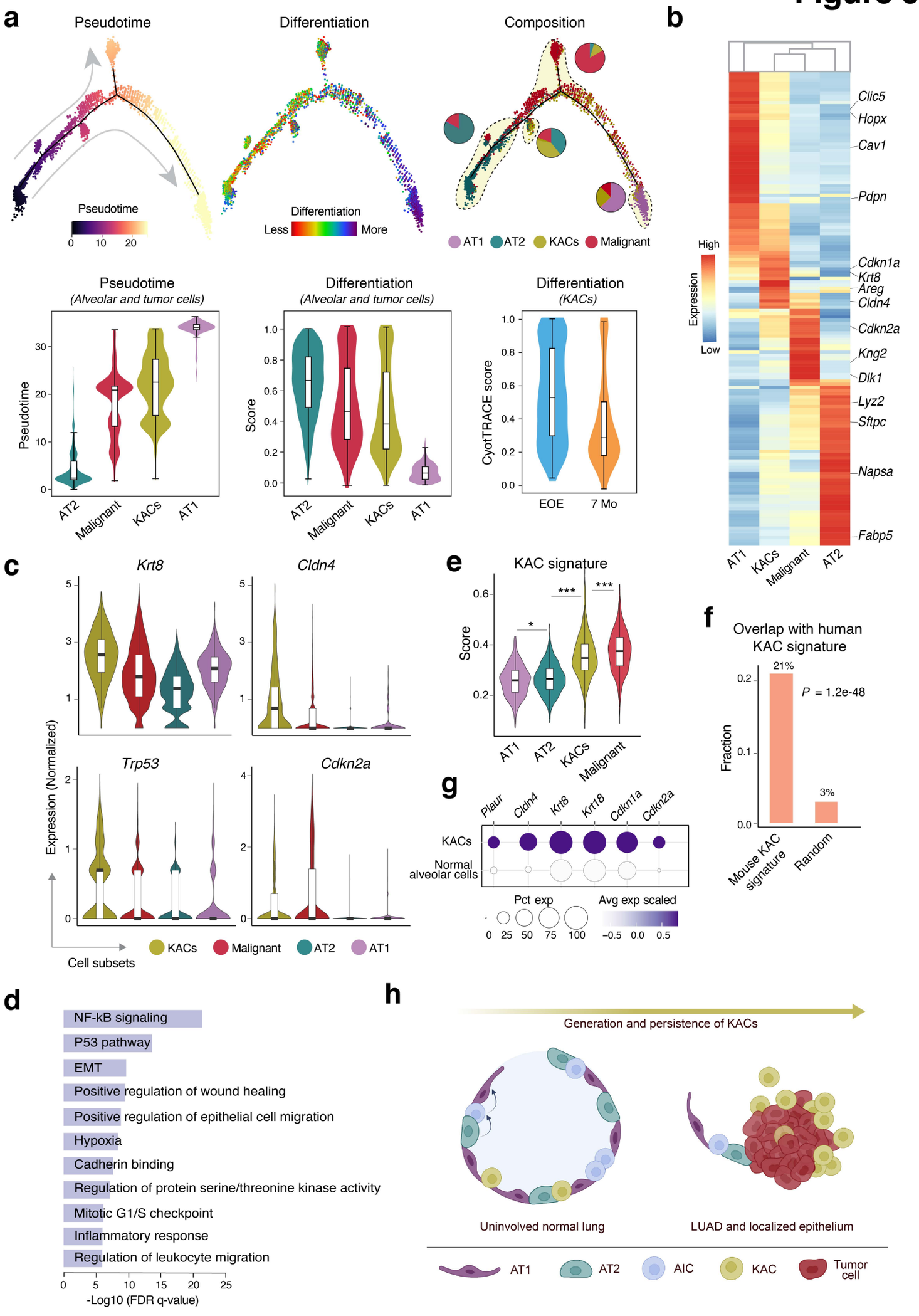

Fig. 5 | *Krt8*+ alveolar cells are implicated in transition of alveolar cells to Kras-mutant malignant cells. a, Developmental trajectory of alveolar and malignant cells colored by inferred pseudotime (top left), cell differentiation status (top middle), and cell type (top right). Distribution of inferred pseudotime (bottom left) and CytoTRACE (bottom middle) across the indicated cell subsets. CytoTRACE score distribution in KACs at the two time points. EOE: end of exposure (bottom right). 7 mo: 7 months post exposure to NNK. Box, median ± interquartile range; whiskers, 1.5× interquartile range; center line: median. **b**, Normalized expression level of differentially expressed genes in AT1, KACs, malignant cells and AT2 cells. **c**, Expression levels of representative KAC marker genes across 4 indicated cell subsets. Box, median ± interquartile range; whiskers, 1.5× interquartile range; center line: median.. **d**, Significantly enriched functional pathways of marker genes of KACs. **e**, Enrichment of KAC signature score, defined using marker genes of KACs, across the four cell subsets. ***: FDR q-value < 0.001; *: FDR q-value < 0.05. P-values were calculated using Mann-Whitney U test. Boxes definitions are the same as in panel b. **f**, Fraction of genes belonging to human KAC signature among mouse KAC genes versus in genome background. *P*-value was calculated using Fisher's exact test. **g**, Proportions and average expression levels of mouse orthologs of selected human KAC marker genes in murine KACs and other alveolar cells. **h**, Proposed model for alveolar plasticity dynamics in lung cancer whereby a subset of the AICs in the intermediate AT2-to-AT1 differentiation state can evolve into KACs and later, lead to LUAD development.