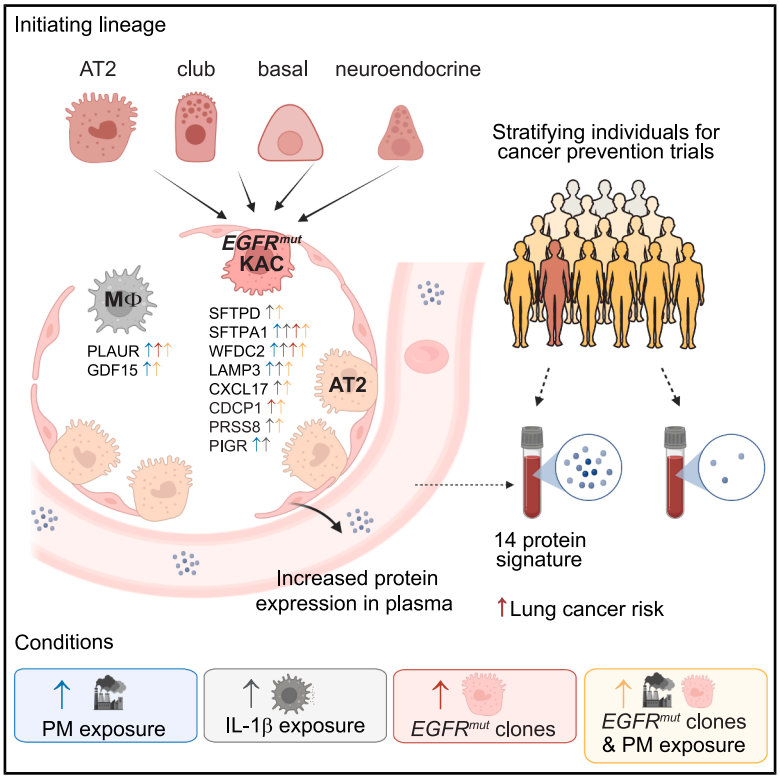
Cell

Plasma signals of lung tumor promotion for molecular cancer prevention

Graphical abstract



Highlights

- A 14-protein plasma signature predicts lung cancer more than 5 years before diagnosis
- Diverse epithelial lineages converge on a transitional state in *EGFR*-driven LUAD
- Particulate matter, oncogenic mutations, and IL-1β elevate the plasma signature
- The signature stratifies benefit from anti-IL-1β lung cancer prevention

Authors

Tej Pandya, Maria Zagorulya, Michelle M. Leung, ..., William Hill, Clare E. Weeden, Charles Swanton

Correspondence

william.hill@cruk.manchester.ac.uk (W.H.),
weeden.c@wehi.edu.au (C.E.W.),
charles.swanton@crick.ac.uk (C.S.)

In brief

A 14-protein plasma signature identifies individuals benefiting from anti-IL-1β-based lung cancer risk reduction and demonstrates how diverse tumor-promoting factors converge on the induction of an alveolar transitional state that underlies lung tumorigenesis.

Pandya et al., 2026, Cell *189*, 1–19
June 25, 2026
https://doi.org/10.1016/j.cell.2026.05.005

CellPress

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

Cell

CellPress
OPEN ACCESS

Article

Plasma signals of lung tumor promotion for molecular cancer prevention

Tej Pandya,[1,2,3,66] Maria Zagorulya,[1,66] Michelle M. Leung,[1,3,4,66] Marcellus Augustine,[1,3,5,66] Lydia Y. Liu,[1,3,67] Aino-Maija Leppä,[1,3,67] Ulysse Baruchel,[1,3,67] Sin Wi Ng,[6,67] Tamara Klockner,[1,67] Miriam Mugabo,[1,3,67] Anthony J. Griffen,[7,8] Oleg Blyuss,[9,10] Chrysante S. Iliakis,[11] Amalie Grenov,[12] Kerstin Haase,[3,13] David C. Muller,[14] Ka Hung Chan,[15,16] Jincheng Wu,[17] Vernon A. Burk,[18] Neil Wright,[15] Alix Le Marois,[19] Ekaterina Pazukhina,[9] Sophia Ward,[1,3,20] Hubert Slawinski,[20] Marc Pelletier,[17] Cian Murphy,[1,3] Matthew D. Park,[21,22,23] Thomas Snoeks,[24] Alejandro Suarez-Bonnet,[25,26] Simon L. Priestnall,[25,26] Alexandros Hardas,[25] Charlotte Grieco,[1,3] Ami Archer,[1] Alpkaan Celik,[1] Alejandro Jimenez-Sanchez,[27] Rachel Scott,[3,13] Hana Zahed,[28] Léa Montégut,[21] Rafael Meza,[29] Clinton H. Durney,[29] Stephen Lam,[30] Takahiro Karasaki,[1,3,31] Roel C.H. Vermeulen,[32] Huilei Xu,[33] Pablo Serrano-Fernandez,[34] Tatjana Crnogorac-Jurcevic,[35] Usha Menon,[36] Sophia Apostolidou,[36] Alexey Zaikin,[37,38]

(Author list continued on next page)

[1]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK
[2]UKRI UCL Centre for Doctoral Training in AI-enabled Healthcare Systems, University College London, London, UK
[3]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK
[4]Cancer Genome Evolution Research Group, University College London Cancer Institute, London, UK
[5]Division of Medicine, University College London, London, UK
[6]Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
[7]Department of Cell Biology, Albert Einstein College of Medicine, Montefiore Health System, Bronx, New York, NY, USA
[8]Montefiore Einstein Comprehensive Cancer Center, Albert Einstein College of Medicine, Bronx, New York, NY, USA
[9]Centre for Cancer Screening, Prevention and Early Diagnosis, Wolfson Institute of Population Health, Queen Mary University of London, London, UK
[10]Department of Paediatrics and Paediatric Infectious Diseases, Institute of Child's Health, I.M. Sechenov First Moscow State Medical University, Moscow, Russia
[11]Immunoregulation Laboratory, The Francis Crick Institute, London, UK
[12]Autoimmunity Laboratory, The Francis Crick Institute, London, UK
[13]Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK
[14]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
[15]Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[16]Oxford Global Health, University of Oxford, Oxford, UK
[17]Novartis, Cambridge, MA, USA

(Affiliations continued on next page)

SUMMARY

Predicting lung cancer risk would enhance prevention trials. Although the Canakinumab Anti-inflammatory Thrombosis Outcome Study (CANTOS) trial demonstrated reduced lung cancer incidence with interleukin (IL)-1β inhibition, the high number needed to treat (NNT) to prevent lung cancer limits its use in unselected populations. Using machine learning, we identified a 14-protein plasma signature predicting lung cancer more than 5 years before diagnosis. The signature, validated across eight cohorts, was elevated in current smokers and individuals exposed to particulate matter (PM) and linked to lung myeloid and alveolar cells. In epidermal growth factor receptor (*EGFR*)-driven lung adenocarcinoma, diverse epithelial lineages converged on a keratin8^{+}/claudin4^{+} alveolar transitional state (KAC), whose transcriptional programs correlated with signature emergence. Components of the signature were induced by PM, oncogenic *EGFR*, or IL-1β, whereas IL-1β inhibition restrained PM-driven KAC expansion and early tumorigenesis. In CANTOS, the signature identified individuals who seemed to benefit more from anti-IL-1β therapy, lowering the NNT threshold and nominating circulating signals of tumor promotion for prevention.

INTRODUCTION

Although lung cancer screening programs in the USA, Europe, and Australia are indicated in those over the age of 50 with significant smoking histories,[1] cancer rates within this population remain too low to enable selection of at-risk individuals for prevention trials.[2] Additionally, these selection criteria do not capture light or never-smoking individuals, populations in which

Richard Gunu,[39] Harry J. Whitwell,[40,41] Zhe Huang,[42] Zonglun Li,[9,39] Xin Hu,[43] Bo Zhu,[43] Liming Li,[44,45,46] María-Dolores Chirlaque,[47,48,49] Marcela Guevara,[48,50,51] P. Martijn Kolijn,[32] Aghiles Guenoun,[28] Neeloffer Mookherjee,[52] Mattias Johansson,[28] Ziqiao Wang,[53] Nilanjan Chatterjee,[53] Chao-Hua Chiu,[54] Zhengming Chen,[15] Dana Pe'er,[27,55] Erik Sahai,[19] Saskia Freytag,[6,56] Andreas Wack,[11] Marc J. Gunter,[14,57] Miriam Merad,[21] Jianjun Zhang,[43] Christopher Carlsten,[58] Pan-Chyr Yang,[59,60,61] Hsuan-Yu Chen,[62] Elizabeth A. Platz,[18] Lindsay M. LaFave,[7,8] Karl Smith-Byrne,[63] Mariam Jamal-Hanjani,[3,13,64] Kevin Litchfield,[3] Nuno R. Nene,[1] Nicholas McGranahan,[3,4] Eva Grönroos,[1] William Hill,[1,65,68,*] Clare E. Weeden,[1,6,56,68,*] and Charles Swanton[1,3,64,68,69,*]

[18]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[19]Tumour Cell Biology Laboratory, The Francis Crick Institute, London, UK
[20]Genomics Science Technology Platform, The Francis Crick Institute, London, UK
[21]Marc and Jennifer Lipschultz Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[22]The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[23]Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[24]Imaging Research Facility, The Francis Crick Institute, London, UK
[25]Department of Pathobiology & Population Sciences, The Royal Veterinary College, London, UK
[26]Experimental Histopathology, The Francis Crick Institute, London, UK
[27]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[28]Early Detection, Prevention and Infections Branch (EPR), International Agency for Research on Cancer (IARC/WHO), Lyon, France
[29]Department of Population Health Sciences, British Columbia Cancer Research Institute, Vancouver, BC, Canada
[30]Department of Basic and Translational Research, British Columbia Cancer Research Institute, Vancouver, BC, Canada
[31]Department of Thoracic Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[32]Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, the Netherlands
[33]Novartis Biomedical Research, Cambridge, MA, USA
[34]Novartis Pharmaceuticals, Basel, Switzerland
[35]Centre for Cancer Biomarkers and Biotherapeutics, Barts Cancer Institute, Queen Mary University of London, London, UK
[36]UCL Innovative Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK
[37]Department of Mathematics and Institute for Women's Health, University College London, London, UK
[38]Centre for Cognition and Decision making, Institute for Cognitive Neuroscience, HSE University, Moscow, Russia
[39]Department of Women's Cancer, Institute for Women's Health, University College London, London, UK
[40]National Phenome Centre and Imperial Clinical Phenotyping Centre, Department of Metabolism, Digestion and Reproduction, IRDB, Imperial College London, London, UK
[41]Section of Bioanalytical Chemistry, Burlington Danes Building, Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK
[42]Nuffield Department of Population Health, University of Oxford, Oxford, UK
[43]Department of Thoracic/Head and Neck Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA
[44]Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China
[45]Peking University Center for Public Health and Epidemic Preparedness & Response, Beijing, China
[46]Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China
[47]Department of Epidemiology, Regional Health Council, Murcia University, Murcia, Spain
[48]CIBER in Epidemiology and Public Health (CIBERESP), Madrid, Spain
[49]Biomedical Research Institute of Murcia Pascual Parrilla-IMIB, 30120 Murcia, Spain
[50]Instituto de Salud Pública y Laboral de Navarra, 31003 Pamplona, Spain
[51]Navarra Institute for Health Research (IdiSNA), 31008 Pamplona, Spain
[52]Max Rady College of Medicine, University of Manitoba, Winnipeg, MB, Canada
[53]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[54]Department of Internal Medicine, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan
[55]Howard Hughes Medical Institute, New York, NY, USA
[56]Department of Medical Biology, University of Melbourne, Melbourne, VIC, Australia
[57]Cancer Epidemiology and Prevention Research Unit, School of Public Health, Imperial College London, London, UK
[58]Air Pollution Exposure Laboratory, Division of Respiratory Medicine, Department of Medicine, Vancouver Coastal Health Research Institute, The University of British Columbia, Vancouver, BC, Canada
[59]National Taiwan University YongLin Institute of Health, National Taiwan University, Taipei 106, Taiwan
[60]Department of Internal Medicine, National Taiwan University College of Medicine, Taipei 100, Taiwan
[61]Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan
[62]Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan
[63]Cancer Epidemiology Unit, University of Oxford, Oxford OX3 7LF UK
[64]Department of Oncology, University College London Hospitals, London, UK
[65]Cancer Research UK Manchester Institute, The University of Manchester, Manchester, UK
[66]These authors contributed equally
[67]These authors contributed equally
[68]Senior author
[69]Lead contact
*Correspondence: william.hill@cruk.manchester.ac.uk (W.H.), weeden.c@wehi.edu.au (C.E.W.), charles.swanton@crick.ac.uk (C.S.)
https://doi.org/10.1016/j.cell.2026.05.005

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

CellPress
OPEN ACCESS

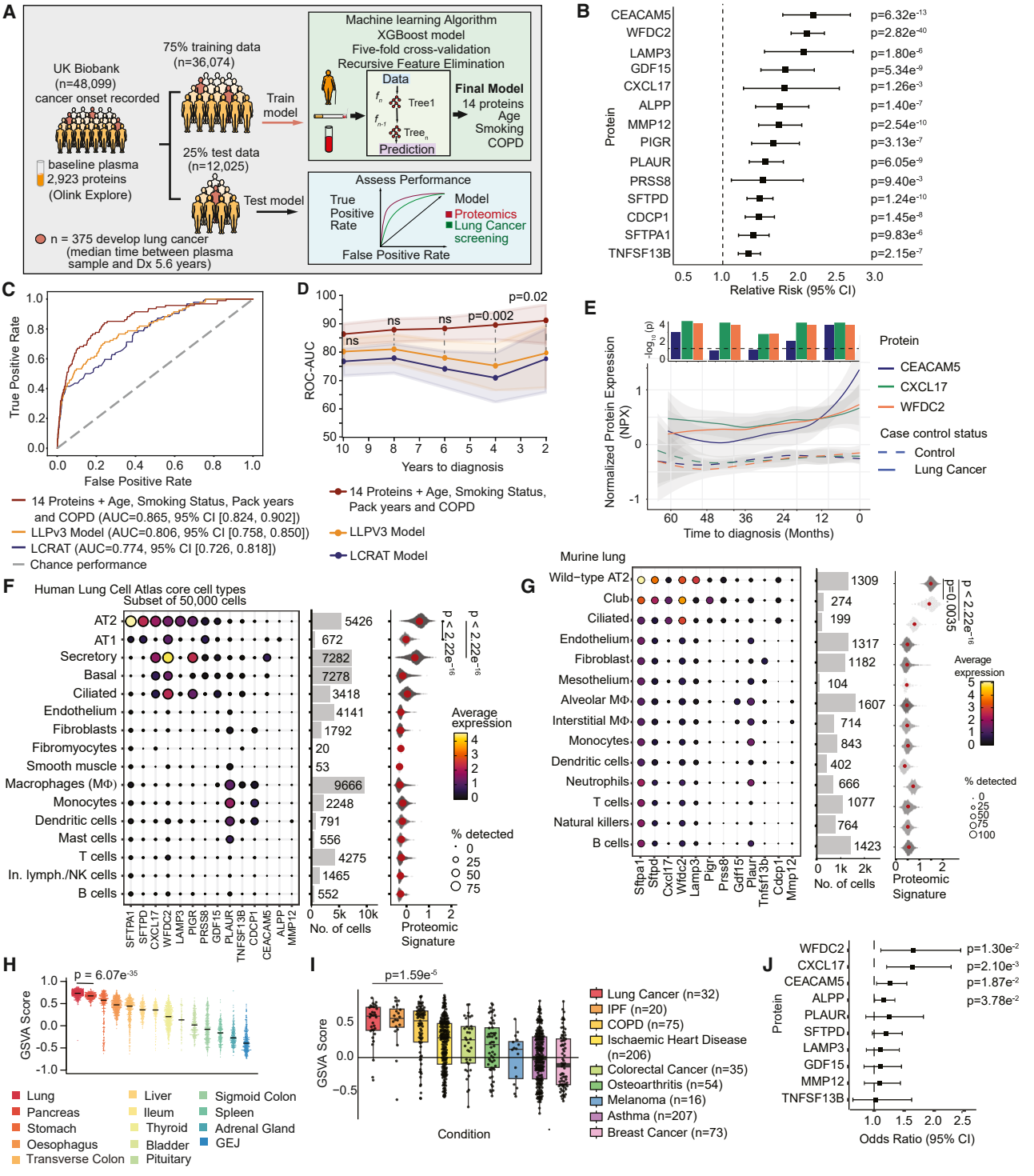


Figure 1. Lung-associated signals of tumorigenesis are present in the blood years before clinical diagnosis

(A) Schematic of the machine learning framework used to identify plasma proteins associated with incident lung cancer diagnosis in the UK Biobank (UKBB). Smoking, smoking status and pack-years. Dx, diagnosis.

(B) Random-effects meta-analysis of the 14 proteins across eight external datasets with relative risk and 95% confidence intervals for each protein (Wald test).

(C) ROC-AUC on the held-out dataset of 12,025 individuals (n = 75 cases) comparing the machine learning model with the LCRAT and LLPv3 models (DeLong's test).

(D) Model performance by time to diagnosis (preceding 2-year intervals; DeLong's test, shaded interval represents 95% CI).

(legend continued on next page)

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

lung adenocarcinoma (LUAD) predominates and activating *EGFR* mutations represent the most common oncogenic driver.[3] Identifying molecular events preceding malignant transformation could aid risk stratification and prevention strategies.

Seemingly normal lung tissue harbors cells with mutations in cancer genes,[4] underscoring that mutations may be necessary but are rarely sufficient for oncogenesis.[5] We have previously reported that particulate matter (PM) exposure triggers macrophages to release interleukin-1 beta (IL-1β), which enhances the LUAD-initiating potential of mutant epithelial cells.[6] The tumor-promoting role of IL-1β is further evidenced by the phase 3 Canakinumab Anti-inflammatory Thrombosis Outcome Study (CANTOS) randomized controlled cardiovascular prevention trial, which reported a dose-dependent reduction in lung cancer incidence following anti-IL-1β therapy.[7] However, therapeutic trials of canakinumab in patients with established non-small cell lung cancer (NSCLC) in the first-line metastatic or adjuvant setting failed to demonstrate efficacy.[8,9] These findings indicate that anti-IL-1β treatment may reduce lung cancer incidence but has no effect on established disease, suggesting a therapeutic window for treatment efficacy that precedes clinically detectable cancers. Biomarkers that identify high-risk individuals prior to malignancy are therefore essential to deliver molecular cancer prevention via IL-1β inhibition. Defining the interplay between tumor-initiating cells, exogenous challenge by tumor promoters, and microenvironmental context may enable better prediction of lung cancer risk.

The lung epithelium comprises airway cells, including basal, club, and neuroendocrine progenitor cells, as well as alveolar type I and type II (AT1 and AT2) cells, of which the latter function as a facultative progenitor upon alveolar injury.[10] Both the initiating epithelial lineage and the nature of the oncogenic driver determine lung cancer subtype.[11] While club, AT1, and AT2 cells can give rise to *Kras*-driven LUAD,[12–15] it is unknown whether other stem and progenitor lineages possess LUAD-initiating capacity, particularly in the context of *EGFR* mutations. *Kras*-mutant AT2 cells proceed through a keratin 8$^+$ (Krt8$^+$) alveolar intermediate cell (KAC) state with enhanced tumor-seeding potential (also termed highly plastic cell state) that is required for malignant progression.[16–19] Oncogenic KACs share transcriptional similarities with damage-associated alveolar epithelial states, which transiently arise from AT2 cells during their transition toward mature AT1 cells following lung injury.[20–24] KACs are therefore emerging as a putative target linking lung injury and repair responses with pathological epithelial cell states in not only lung cancer but also pulmonary fibrosis and chronic obstructive pulmonary disease (COPD).[25,26]

Here, we applied a machine learning framework to population-scale UK Biobank proteomic data and identified a 14-protein plasma signature, predictive of future lung cancer risk, that replicated in eight external datasets. We addressed the origins of the signature, first determining that basal, club, neuroendocrine, and AT2 epithelial lineages were competent to form *EGFR*-driven LUAD, with conserved KAC-like states observed within the alveolar niche. Consistent with the 14-protein signature reflecting a tumor-promoted niche, the expression of the signature increased in myeloid and wild-type AT2 cells in the lung upon PM exposure, in the presence of *EGFR*-mutant clones, or following IL-1β exposure. Blocking IL-1β signaling limited the outgrowth and potency of KACs, highlighting their role as an early lung cancer prevention target. In the CANTOS proteomic sub-cohort, a high protein signature at trial entry identified individuals with a reduction in lung cancer incidence following anti-IL-1β therapy, supporting the use of a plasma-based signature for patient selection in lung cancer prevention.

RESULTS

Lung-associated signals of tumorigenesis are present in the blood years before clinical diagnosis

We developed a machine learning framework to predict incident lung cancer diagnoses using data from the population-level UK Biobank Pharma Proteomics Project, integrating plasma proteomic profiles (n = 2,923 proteins) collected at baseline with subsequent cancer registry outcomes (median 5.6 years to lung cancer diagnosis, range 0.16–11.00 years)[27] (Figure 1A). We implemented a 75:25 train-test split (N = 48,099 individuals, N = 375 lung cancer cases), utilizing recursive feature elimination to identify a parsimonious predictive model of incident lung cancer cases (Figure S1A). From this approach, we identified 14 proteins together with four patient characteristics (age, smoking status, pack years, and past diagnosis of COPD) that were linked with future lung cancer incidence (Figure 1A; STAR Methods). The 14 proteins are associated with inflammatory signaling (CXCL17,[28] CDCP1,[29,30] GDF15,[31] PIGR,[32] TNFSF13B,[33] and PLAUR[34]), extracellular matrix remodeling (MMP12[35]), epithelial secretion or shedding (CEACAM5,[36] WFDC2,[37] ALPP,[38] and PRSS8[39]), and pulmonary surfactant production (LAMP3,[40] SFTPD, and SFTPA1[41]). We examined the association of these 14 proteins with lung cancer incidence in eight additional proteomic datasets from the UK, US, Iceland, China, and two multinational cohorts, comprising 2,198 incident lung cancer cases and 53,641 non-cancer controls in total (plasma collected a

(E) Locally estimated scatterplot smoothing (LOESS) curves of three proteins measured longitudinally at 1-year intervals (98 cases, 150 controls, median 5 samples per individual) from the UKCTOCS clinical trial.[42] Bars indicate the significance of case-control differences at yearly intervals (Wilcoxon test, capped at $-\log_{10} p = 4$).

(F and G) Single-cell expression of genes encoding the signature proteins, their aggregated score and the number of cells analyzed (F) in the Human Lung Cell Atlas[49] (50,000 cell subset) and (G) mouse lung microenvironment scRNA-seq dataset, enriched for epithelial and other stromal cell types (Wilcoxon test).

(H) GSVA score of the 14-protein signature across tissues analyzed from 19,788 samples (946 donors) in the GTEx Consortium[50] (Wilcoxon test). First significant p value shown.

(I) Boxplot of GSVA scores with interquartile range across individuals with incident disease within 5 years of sampling in UKBB (held-out set; Wilcoxon test). First significant p value shown.

(J) Odds ratios with 95% CI for association with lung cancer for 10 proteins in the TALENT study (n = 251 cases, 501 controls; p values shown for $p < 0.05$ by Wald test).

See also Figures S1 and S2 and Tables S1, S2, S3, and S4.

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

median 7.55 years before lung cancer diagnosis with a range of medians 1.60–12.62 years; Table S1).[42–48] All 14 proteins positively associated with future lung cancer incidence across a random-effects meta-analysis (Figures 1B, S1B, and S1C). We found no significant differences in risk estimates between histological subtypes of lung cancer for each protein in the three cohorts where data were available (Figures S1D–S1F).

To evaluate the ability of the machine learning model to predict incident lung cancer diagnosis, we compared the performance of the model against commonly used lung cancer screening models,[51] using the held-out UK Biobank test data (total N = 12,025, lung cancer N = 75, median time to lung cancer diagnosis 5.1 years, range 0.16–11.00 years; STAR Methods). Among these screening models, the Liverpool Lung Project version 3 model (LLPv3)[52] had the highest receiver operating characteristic-area under the curve (ROC-AUC) on this dataset (Table S2). The machine learning model outperformed the LLPv3 model (p = 0.01 by DeLong's test; Figure 1C), achieving a sensitivity of 0.776 (95% confidence interval [CI]: 0.687–0.857), compared with 0.622 (95% CI: 0.518–0.718) for the LLPv3 model (p = 0.0012 by DeLong's test comparing sensitivities at a fixed false-positive rate of 20%). The greatest improvement in performance of the machine learning model compared with both published lung cancer screening models was observed 2–4 years prior to diagnosis (p = 0.002 by DeLong's test of the machine learning model compared with LLPv3; Figure 1D). Comparing XGBoost models built on the 14 proteins alone, patient characteristics alone, or both, we found protein- and clinical-only models performed comparably (DeLong's test, p = 0.26), while the combined model significantly outperformed each (Figure S2A). Proteomics data for 3/14 proteins (WFDC2, CXCL17, and CEACAM5) were available annually for 5 years preceding lung cancer diagnosis in 98 cases and 150 controls from the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS).[42] Independently of smoking status, these three proteins increased in individuals who developed lung cancer, compared with controls, 2 years before diagnosis (Figures 1E and S2B). Together, these findings suggest that the 14-protein signature augments the ability to predict the future risk of lung cancer when combined with patient characteristics.

To investigate which organs and cell types express the genes encoded in the machine learning-derived 14-protein signature, we interrogated bulk RNA sequencing (RNA-seq) data from healthy human tissues.[50] Transcripts encoding these proteins were more abundant in the human lung compared with other tissues ($p = 6.07 \times 10^{-35}$, Wilcoxon test; Figures 1H and S2C). Analysis of single-cell RNA-seq (scRNA-seq) data from the Human Lung Cell Atlas[49] demonstrated that expression of genes encoding these 14 proteins was predominantly enriched in AT2 cells followed by secretory airway epithelial cells ($p < 2.2 \times 10^{-16}$ between signature score of AT2 cells and secretory cells; Figure 1F), with a subset of genes (*PLAUR*, *TNFSF13B*, *CDCP1*, and *MMP12*) upregulated in myeloid cells, (macrophages, monocytes, and dendritic cells) and fibroblasts (Table S3). We generated a scRNA-seq dataset of immune, epithelial, and all other cell types from the lungs of healthy control mice (see STAR Methods and Figure 1G), capturing 12/14 genes of interest. These genes were similarly enriched in wild-type AT2 and club epithelial cells, with expression of *Plaur*, *Gdf15*, and *Mmp12* largely restricted to myeloid cells and *Tnfsf13b* to fibroblasts in mice (Table S4). We conclude that the circulating 14-protein plasma signature is enriched for genes expressed in wild-type lung epithelial, myeloid, and fibroblast cells in both mice and humans.

To assess specificity, we applied gene set variation analysis (GSVA) to evaluate enrichment across a selection of other diseases within 5 years of baseline plasma sampling in the UK Biobank held-out set. We found the 14-protein signature was higher in respiratory diseases such as idiopathic pulmonary fibrosis (IPF) and COPD ($p = 1.59 \times 10^{-5}$ Wilcoxon test, lung cancer vs. ischemic heart disease; Figure 1I). To understand changes during malignant progression, we performed GSVA of plasma proteomics in the TRACERx (tracking cancer evolution through therapy [Rx]) cohort study,[53] with baseline samples taken prior to resection and at a follow-up from individuals who had not relapsed at least 2 years post-surgery. We found that the 14-protein signature was not elevated in patients with higher stage tumors in TRACERx baseline samples and did not change following surgical resection of the primary tumor, suggesting that the protein signature does not arise from established lung cancer (Figures S2D and S2E).

Given that never-smoker lung cancer cases were scarce in the UK cohorts, we interrogated the association between the protein signature and lung cancer in a predominantly (93.3%) never-smoker population from individuals enrolled in the multi-center, prospective TALENT (Taiwan Lung Cancer Screening in Never-Smoker Trial) clinical trial.[54] We performed Olink proteomics on baseline plasma samples taken from a subset of TALENT participants (N = 251 cases and N = 501 age-, sex-, and baseline smoking status-matched controls; 81.3% female, median time to diagnosis 144 days, range 9–3,492 days, 87.6% stage IA disease, and 62.1% adenocarcinomas). Four proteins (WFDC2, CXCL17, CEACAM5, and ALPP), as well as the overall signature, were associated with future lung cancer diagnoses in TALENT (Figures 1J and S2F), with three of these proteins (CXCL17, CEACAM5, and WFDC2) also increased in the plasma proteome in never-smokers in UKCTOCS prior to lung cancer diagnosis (Figure S2B). We conclude that the 14-protein signature identifies individuals at increased risk of developing lung cancer, fibrotic lung disease, and COPD. In addition, a subset of these proteins may also have utility for risk stratification in never-smoker populations.

Epithelial lineages converge upon KAC states en route to LUAD

Given that an ideal risk signal would encompass all routes to lung cancer initiation, we used *EGFR*-mutant mouse models to probe the origins of the 14-protein signature in early LUAD tumorigenesis. We therefore established which cells form LUAD upon acquiring an *EGFR* mutation. Lineage-restricted Cre adenoviruses[13,55,56] were delivered intratracheally to activate a reporter allele (T mice; $Rosa26^{LSL\text{-}tdTomato/+}$), reporter plus $EGFR^{L858R}$ mutation (ET; $Rosa26^{LSL\text{-}tTa/LSL\text{-}tdTomato}$; $TetO\text{-}EGFR^{L858R}$), or reporter, $EGFR^{L858R}$, and Trp53 loss (EPT; $Rosa26^{LSL\text{-}tTa/LSL\text{-}tdTomato}$; $TetO\text{-}EGFR^{L858R}$; $Trp53^{fl/fl}$) in basal (Ad5-bk5-Cre, requiring polidocanol pre-treatment[56]),

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

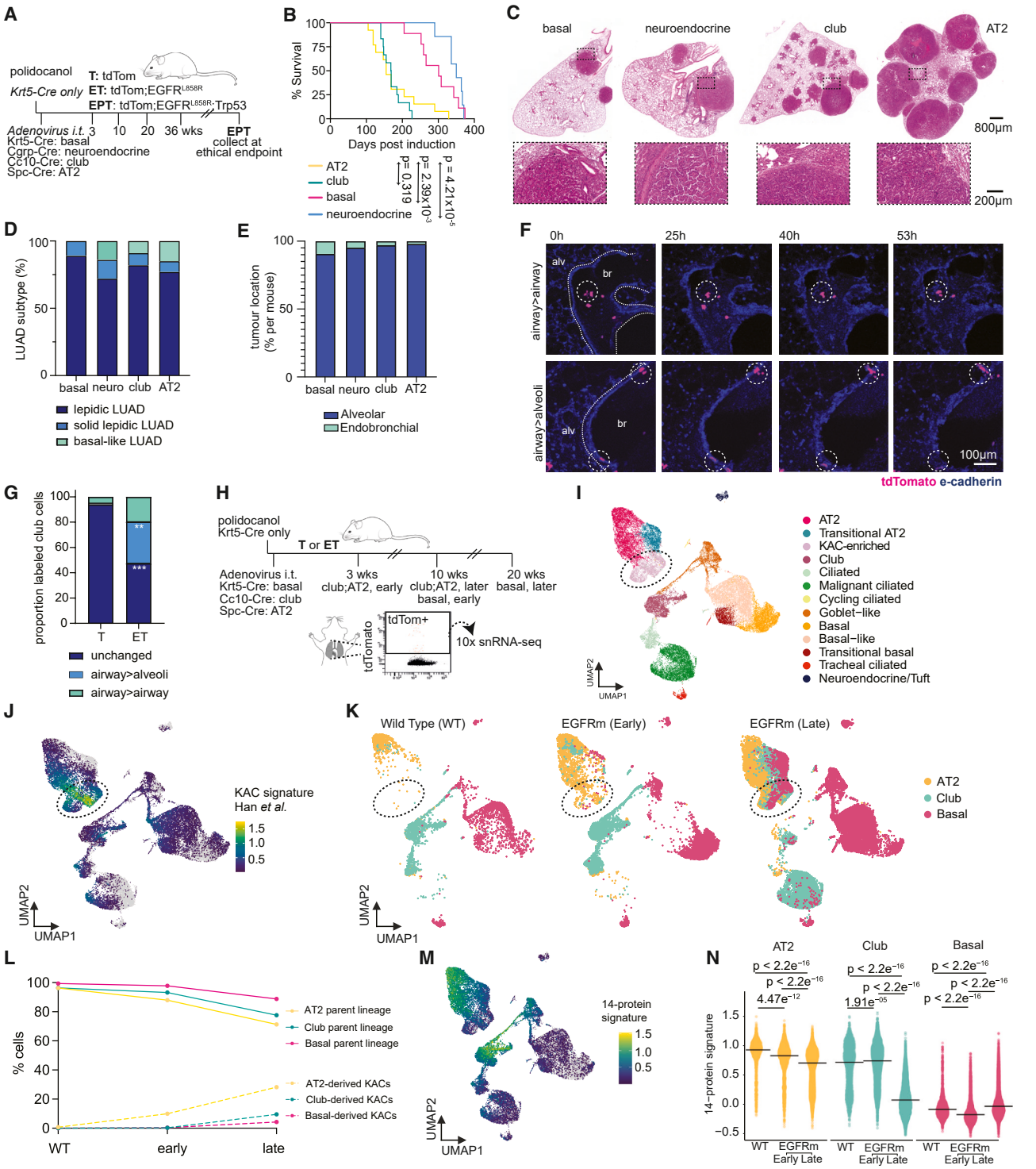


Figure 2. Epithelial lineages converge upon KAC states en route to LUAD

(A) Experimental design of lineage-restricted activation of T, ET, or EPT alleles *in vivo*. EPT mice were monitored monthly by microCT imaging and collected at the indicated time points.

(B–E) Same cohorts throughout. (B) Survival of EPT mice induced with lineage-restricted viruses (two independent cohorts; mice culled for non-cancer illness excluded). $n = 9$ basal, 7 neuroendocrine, 12 club, 13 AT2; log-rank test. (C) Representative histology at the ethical endpoint. (D) Histopathology grade of the most advanced lesion per mouse. (Neuro, neuroendocrine). (E) Spatial tumor location at endpoint. (D and E) Non-significant, two-way ANOVA, Sidak's correction.

(legend continued on next page)

neuroendocrine (Ad5-CGRP-Cre), club (Ad5-CC10-Cre), or AT2 cells (Ad5-SPC-Cre; Figure 2A). In T mice, recombination was rare (basal-targeted: 1 tdTomato$^+$ cell per 1,080 cells analyzed; neuroendocrine: 1/11,800; club: 1/6,400; and AT2: 1/23,600; Figures S3A and S3B) and restricted to the expected epithelial compartments (Figures S3C–S3E; Table S5). In polidocanol-treated basal-targeted mice, recombination occurred in the trachea and extra- and intrapulmonary bronchi, closely recapitulating human basal cell distribution (Figures S3C–S3E; Table S5). These tools enable *EGFR* mutation alone to model early tumor initiation, or with Trp53 loss, across four lineages at low clonal density.

Lineage-specific oncogenic competence was assessed in the more aggressive EPT model[57] to maximize adenocarcinoma transformation within the mouse lifespan. Polidocanol did not impact survival after ubiquitous Ad5-CMV-Cre induction, supporting its use in basal-targeted conditions (Figure S3F). All four lineages formed lung nodules detectable by micro-computed tomography (microCT), with club- and AT2-induced conditions exhibiting shorter latencies and reduced survival in EPT mice (Figures 2B and S3G). At the endpoint, each lineage formed EGFR^{L858R+} LUADs indistinguishable by histology and bulk transcriptomics (Figures 2C, 2D, S3H, and S3I). Surfactant protein C (SPC) protein, an AT2 marker, was ubiquitously expressed across tumors rather than lineage-of-origin-specific markers (Figures S3J and S3K). Temporal grading in ET and EPT mice showed evolution from mutant clones to alveolar hyperplasia, adenoma, and adenocarcinoma, with slower progression of ET relative to EPT tumors (Figures S3L–S3O). Basal and neuroendocrine lineages advanced more slowly to adenocarcinoma than club and AT2 lineages in EPT mice (Figure S3O). Together, these data support that oncogenic *EGFR* confers tumor-initiating potential across epithelial lineages in concert with *Trp53* deficiency.

Regardless of origin, all lineages formed SPC$^+$ adenocarcinoma in EPT mice within the alveolar microenvironment (Figures 2E, S3J, S4A, and S4B; Videos S1, S2, S3, and S4), consistent with the location of human LUAD.[58] In basal-targeted EPT mice, EGFR^{L858R+} airway cells persisted in the trachea for 12 months without forming lesions and remained SPC$^-$, whereas SPC$^+$ LUADs in the same animal arose exclusively within alveoli (Figures S4C and S4D), suggesting an obligate role for the alveolar compartment in *EGFR*-driven LUAD formation. To determine how mutant airway cells access this niche, we performed live imaging of precision-cut lung slices (PCLSs) generated from club- or basal-targeted ET mice. *EGFR*-mutant club cells migrated to alveoli more frequently than tdTomato$^+$ controls (33% vs. 1%; Figures 2F, 2G, and S4E; Videos S5, S6, and S7). Rare labeled wild-type and *EGFR*-mutant basal cells in intrapulmonary airways also transitioned to alveoli after polidocanol injury, but events were too infrequent for quantification and may be injury- or mutation-associated (Videos S8 and S9). Inhibition of Wnt signaling, which governs aberrant airway-to-alveolar differentiation during injury,[20,59,60] reduced *EGFR*-mutant club cell migration and SPC expression in the alveoli (Figures S4F–S4H; Videos S10 and S11). These data indicate that *EGFR*-driven LUAD can arise from multiple lineages and are associated with entry into the alveolar compartment.

We next tracked gene expression dynamics in mutant cells throughout LUAD progression to explore relationships with the 14-protein signature. We performed single-nucleus RNA-seq (snRNA-seq) on tdTomato$^+$ *EGFR*-mutant cells induced in basal, club, or AT2 lineages from ET mice at early clonal expansion and later hyperplastic stages, comparing them to lineage-matched *EGFR*-wild-type cells from T mice (37,627 nuclei from 100 mice; Figure 2H). The neuroendocrine lineage was excluded due to cell scarcity. Using published cell-identity gene sets,[17,20] we identified tracheal (including polidocanol injury-induced subsets), bronchi/bronchiolar, and AT2-derived populations, including KACs, a transient Krt8$^+$ alveolar intermediate state with high tumor-seeding potential[17] (Figures 2I, 2J, and S5A; Table S6). This lineage-tagged dataset enables tracing the earliest cell-intrinsic transcriptomic changes as distinct lineages transition to malignant fates following *EGFR* mutation.

In T mice acutely post-recombination, AT2, club, and basal-derived cells formed distinct clusters, confirming virus specificity (Figures 2K left and 2I). At the equivalent time point in ET mice, rare basal- and club-derived cells acquired alveolar-associated transcriptional features not observed under wild-type conditions, with greater proportions of mutant cells from all lineages subsequently converging on KAC-like states during malignant

(F) Single frames isolated from live-cell imaging, depicting 3D reconstruction of PCLS from ET mice induced with a club-targeted virus and collected 6 weeks after oncogene induction, stained with E-cadherin (blue) and detecting endogenous tdTomato expression (pink). Circles indicate *EGFR*-mutant club cells dividing within the airway epithelium (top, coronal view) or transitioning from bronchiolar (br) to alveolar (alv) microenvironment (bottom, transverse view), with bronchiolar structure indicated in the dotted line. Refer to Videos S5, S6, and S7 for full experimental results.

(G) Proportion of T or ET club cells transitioning to the alveolar niche at 6 weeks post-induction (one representative experiment shown; $n = 4$ mice, 2–3 fields of view per mouse; experiment performed twice with consistent results). Two-way ANOVA with Tukey's correction, $^{**}p = 0.0022$, $^{***}p < 0.0001$.

(H) Experimental design to assess basal, club, and AT2 cell fate in early *EGFR*-driven tumorigenesis by isolating lineage-restricted tdTomato+ wild-type (WT) or mutant cells by flow cytometry for snRNA-seq analysis. $n = 10$ animals per genotype/time point/lineage pooled.

(I–K) Uniform manifold approximation and projection (UMAP) visualization of snRNA-seq data from tdTomato$^+$ lineage-traced cells in WT, early (3 weeks in AT2- and club-targeted mice; 10 weeks in basal-targeted mice) and late (10 weeks in AT2- and club-targeted mice; 20 weeks in basal-targeted mice) in ET tumorigenesis colored by (I) cell cluster, (J) murine KAC score from Han et al.,[17] or (K) split by time point and colored by lineage-of-origin. The dotted circle indicates the KAC-enriched cluster. $n = 37,627$ nuclei.

(L) Quantification of the proportion of cells within their parent lineage (defined as the cell clusters present at >1.5% frequency in WT conditions) over time and within the KAC-enriched cluster.

(M and N) UMAP visualization of all nuclei analyzed, colored by the 14-protein gene set signature score and quantified in (N) throughout early tumorigenesis, p values from t tests with Bonferroni's correction.

See also Figures S3, S4, and S5, Tables S5 and S6, and Videos S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, and S11.

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

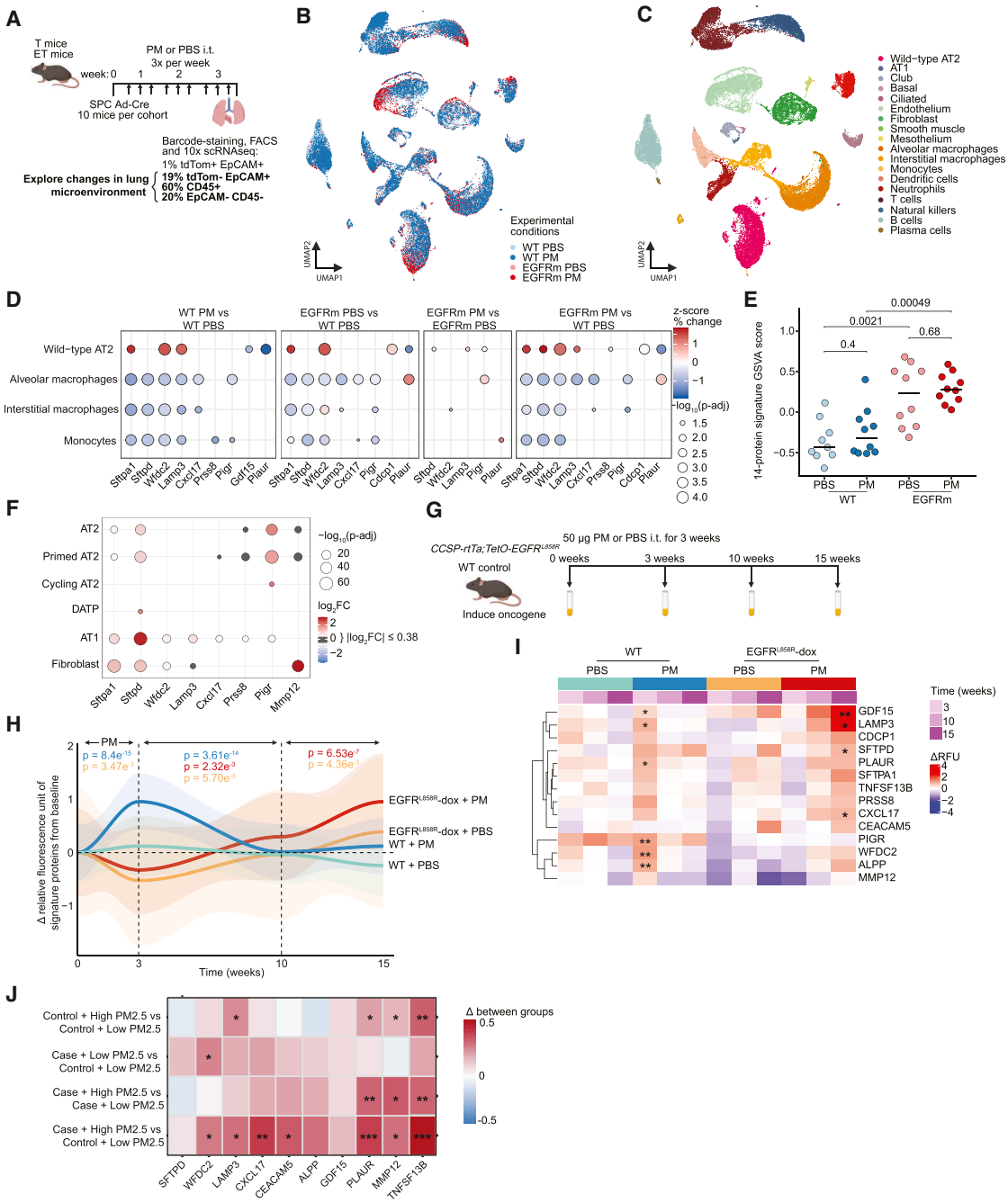


Figure 3. Tumor-promoting inflammatory challenge by air pollution provokes expression of the 14-protein signature

(A) Experimental schematic depicting PBS- or PM-exposed T and ET mouse models of *EGFR*-driven lung tumorigenesis. Lungs were harvested at the 3-week time point, and tdTomato-negative cells were analyzed by scRNA-seq. n = 10 animals per experimental condition.

(B and C) UMAP visualization of lung microenvironment cells colored by (B) experimental condition and (C) cell type.

(D) Bubble plot of the 12 detected transcripts across experimental conditions in WT AT2 cells, alveolar macrophages, interstitial macrophages, and monocytes. Color denotes normalized percentage change; size represents $-\log_{10}$ adjusted p value (Wilcoxon test), where only results of $p < 0.05$ are shown.

(E) GSVA score of the 12 detected transcripts across the four conditions, pseudobulked per mouse, Wilcoxon test.

(F) scRNA-seq data from Choi et al.[21] of mouse AT2 organoids co-cultured with stromal cells, treated with or without IL-1β. Differential expression shown for 12 detected transcripts. Only genes with an adjusted p (Wilcoxon test) < 0.05 are plotted. Color denotes $\log_2$-fold-change; size represents $-\log_{10} p$ value.

(G) Schematic of PM- or PBS-exposed, doxycycline-induced *EGFR*-mutant (CCSP-rtTa; TetO-*EGFR*^{L858R}) or WT C57B/L6 mice. Plasma collected longitudinally for SomaScan proteomics.

(legend continued on next page)

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

progression, reflecting tumorigenesis prior to adenoma formation (Figures 2K, 2L, and S5B). KAC-like states were characterized by common expression of *Cldn4*, *Krt8*, and *Itga2*, while basal- and club-derived KACs maintained elevated lineage-of-origin gene expression (club: *Scgb1a1* and *Alcam*; basal: *Cadm1* and *Sox5*; Figures S5C–S5E). Spatial analyses 20–36 weeks post-induction revealed $EGFR^{L858R+}$ $Cldn4^{+}$ cells in alveolar hyperplasias and adenomas across basal, neuroendocrine, club, and AT2-induced ET mice (Figures S5F and S5G).

Next, we examined epithelial cell-intrinsic expression of the 14-protein signature (12/14 genes captured) and found it highest in lineage-traced club- and AT2-derived cells from wild-type reporter mice but decreasing in $EGFR^{L858R+}$-mutant cells and KACs during tumorigenesis (Figures 2M, 2N, and S5H). Thus, although diverse epithelial lineages converge on KAC-like states during *EGFR*-driven tumorigenesis, the 14-protein signature expression declines within mutant cells as they progress toward malignancy. This observation is consistent with a tumor-extrinsic origin of the signature (Figures S2D and S2E). As we captured only lineage-tagged epithelial populations, we next investigated whether the circulating 14-protein signature instead arises from lung microenvironmental cells prior to malignant transformation.

Tumor-promoting inflammatory challenges by air pollution provoke expression of the 14-protein signature

To elucidate the effect of *EGFR*-mutant clones and PM exposure on the 14-protein signature expression across lung cell types, we performed scRNA-seq on lungs from T and ET mice exposed to PBS or PM (see STAR Methods). We focused our analysis on all 42,463 sequenced *EGFR*-wild-type lung cells across 39 mice and 18 cell types, capturing epithelial- and myeloid-associated signature genes (Figures 3A–3C, S6A, and S6B; Tables S4 and S7). At the 3-week time point, PM exposure did not expand *EGFR*-mutant clones but increased $CD68^{+}$ macrophages[6] with enrichment of IL-1β+ macrophages in alveoli compared with peri-airway regions (Figure S6C). PM exposure alone increased expression of epithelial-associated signature genes in wild-type AT2 cells compared with PBS control (*Sftpa1*, *Wfdc2*, and *Lamp3*; Figures 3D and S6D). By contrast, the presence of *EGFR*-mutant clones, even in the absence of PM, increased transcripts of epithelial (*Wfdc2*, *Cdcp1*, and *Sftpa1*) and myeloid (*Plaur*) signature genes in wild-type cells (Figures 3D and S6D). The combined presence of *EGFR*-mutant clones and PM exposure resulted in wild-type AT2 cells upregulating *Sftpa1*, *Sftpd*, *Lamp3*, *Cdcp1*, and *Wfdc2*, and alveolar macrophages increasing *Plaur* expression (Figures 3D and S6D). Signature enrichment increased in wild-type AT2 cells after PM exposure in the presence of *EGFR*-mutant clones or a combination of both (Figures S6E–S6H), as well as in pseudobulked analysis of the whole-lung tissue in the presence of *EGFR*-mutant clones (Figure 3E). The lack of enrichment with PM alone in wild-type whole-lung analysis may reflect dilution of the epithelial transcriptional signal due to the contribution of non-epithelial cell types. Together these data demonstrate that epithelial- and myeloid-associated components of the signature are altered by the presence of oncogenic *EGFR*-mutant clones, PM or a combination of both.

Given the central role of IL-1β in the pulmonary inflammatory response to PM,[6] we investigated if components of this signature might be induced by IL-1β. Analyses of scRNA-seq data from published mouse AT2 and fibroblast organoid co-cultures treated with IL-1β peptide[21] demonstrated increased transcription of 8/12 genes captured within this system spanning wild-type epithelial cells (*Sftpa1*, *Sftpd*, *Wfdc2*, *Lamp3*, *Cxcl17*, *Prss8*, and *Pigr*) and fibroblast-associated components (*Mmp12*) relative to PBS-treated controls (Figure 3F). Concordant with these data, we found that IL-1β treatment of human wild-type fetal-lung-derived AT2 organoids was sufficient to increase expression of three out of four genes highly expressed in AT2 cells (*WFDC2*, *LAMP3*, and *CXCL17*; Figure S7A). These data indicate that IL-1β is sufficient to drive transcription of an epithelial-associated subset of the 14-protein signature. We further validated one epithelial-associated gene at the protein level, finding that PM treatment of PCLS from wild-type and *EGFR*-mutant murine lungs triggered the release of the Lamp3 protein into the culture supernatant (Figure S7B), suggesting that PM leads to increased release of components of the 14-protein signature into circulation.

Components of the 14-protein signature increase in the plasma of mice and humans in response to environmental exposures

We next investigated whether the observed transcriptional changes in the murine lung resulted in a detectable 14-protein signature in plasma. We used a more aggressive model of $EGFR^{L858R}$-driven LUAD (*EGFR*-dox: *CCSP-rtTa; TetO-EGFR*^{L858R}, with doxycycline-induced expression of $EGFR^{L858R}$ from airway and alveolar lineages; see STAR Methods), where we have previously shown that PM exposure increases the number of LUADs formed at 10 weeks post-oncogene induction.[6] We exposed *EGFR*-dox and wild-type control mice treated with doxycycline to PM or PBS for 3 weeks and collected longitudinal blood samples at baseline, 3 weeks (widespread hyperplasia, rare LUAD present),[61] 10 weeks (widespread hyperplasia with multiple focal LUAD present),[61] and at the ethical endpoint (15 weeks; Figure 3G). In wild-type mice, PM exposure over 3 weeks transiently increased the plasma protein signature, which returned to baseline at 10 weeks (7 weeks post-PM cessation), with increases in both epithelial-associated (Wfdc2, Pigr, and Lamp3) and myeloid-associated proteins (Gdf15, Plaur; Figures 3H and 3I). In contrast, both PBS- and PM-exposed

(H) LOESS curves of the 14-protein signature (median Δ from baseline, $\log_{10}$-transformed and *Z* scored) at 3, 10, and 15 weeks. Shaded areas denote 95% CI (n = 4–6 per group); significance from linear mixed-effects models.

(I) Heatmap of protein levels at 3, 10, and 15 weeks in *EGFR*-mutant and control mice compared with baseline levels (*$p < 0.05$, **$p < 0.01$, significance from linear mixed-effects models).

(J) Change in expression of the 10 available proteins in the TALENT study between individuals who developed lung cancer and controls, split by median $PM_{2.5}$ levels. (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ Wilcoxon test).

See also Figures S6 and S7 and Tables S4, S7, and S8.

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

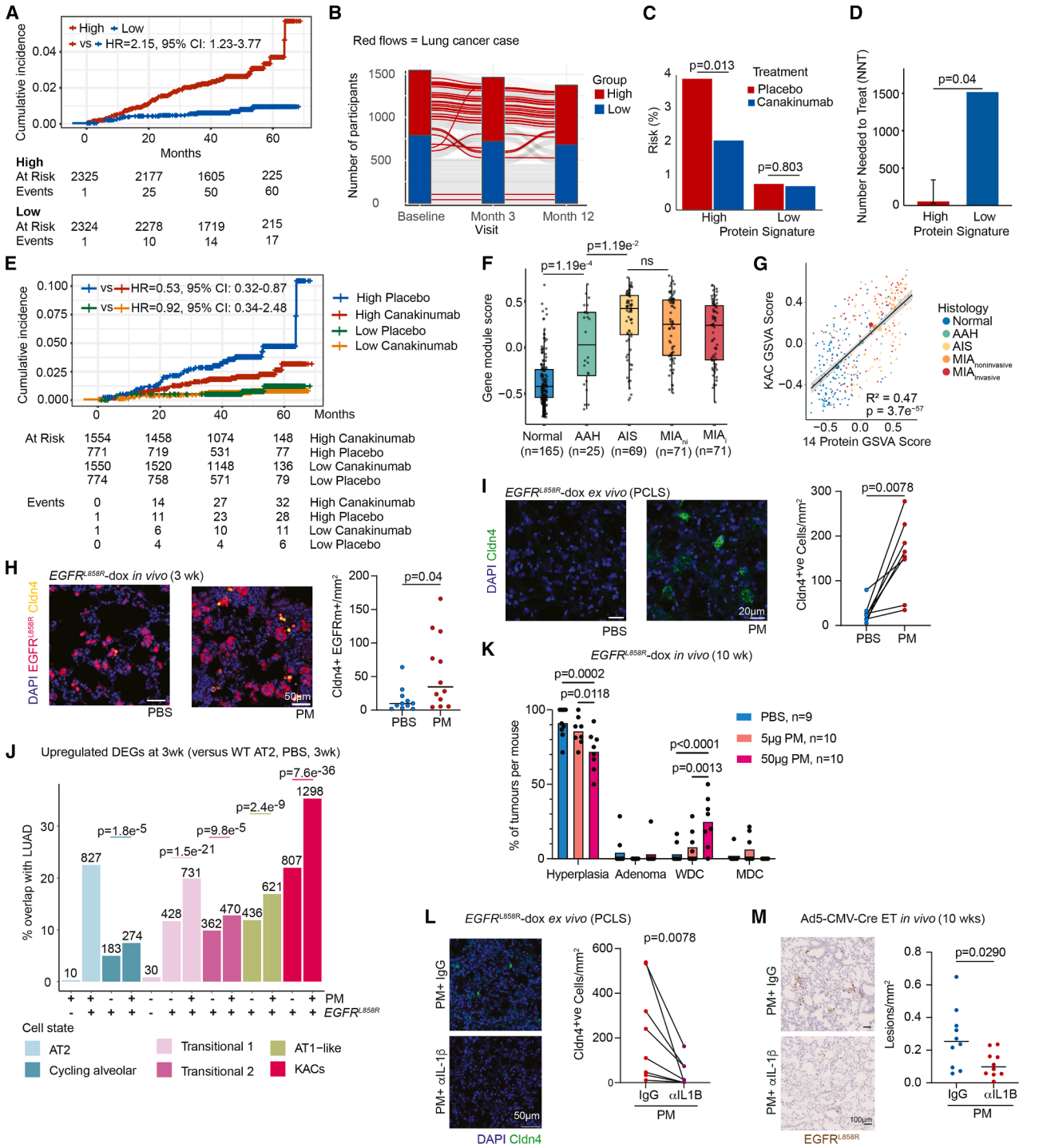


Figure 4. The 14-protein signature stratifies individuals for lung cancer prevention via anti-IL-1β therapy

(A) Cumulative lung cancer incidence in the sub-cohort of the CANTOS trial,[64] stratified by baseline signature levels ($p = 0.002$ by Wald test). Events correspond to the time interval, capped at $n = 60$ months.

(B) Sankey plot showing signature category transitions among placebo-treated individuals from baseline to months 3 and 12. Red flows indicate incident lung cancer.

(C) Lung cancer risk at trial entry in CANTOS by signature level and treatment (chi-squared test).

(D) Number needed to treat (NNT) by signature group (Wald test).

(E) Cumulative lung cancer incidence stratified by signature level and treatment (log-rank test, $p < 0.05$; significant in the high-signature group).

(F) GSVA of the 14-protein signature in pre-invasive lesions and adjacent normal tissue (bulk RNA-seq from Chen et al.[65]). AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma *in situ*; MIA, minimally invasive adenocarcinoma (ni, non-invasive; i, invasive), Dunn test.

(legend continued on next page)

EGFR-dox mice exhibited significant elevation of the signature between 3 and 10 weeks ($p = 2.32 \times 10^{-3}$ for the *EGFR*-PM group and $p = 5.70 \times 10^{-3}$ for the *EGFR*-PBS group; Figure 3H). Further investigation of the components of the protein signature revealed an increase of epithelial- (Lamp3, Sftpd, and Cxcl17) and myeloid-associated (Gdf15) proteins in the plasma in PM-exposed *EGFR*-dox mice (Figure 3I). In summary, these data suggest that the transcriptional changes observed acutely after PM exposure (Figures 3D, 3E, and S6E–S6H) are reflected in transiently elevated 14-protein signatures in wild-type mouse plasma. Sustained elevation of the 14-protein signature is present in plasma as a result of the combination of both *EGFR*-driven initiation and PM-driven tumor promotion (Figure 3H).

We set out to decipher how environmental exposures and cancer risk influence the 14-protein signature in human plasma. In the UK Biobank held-out set, all 14 proteins were associated with ever-smokers compared with never-smokers (Figure S7C), with the mean signature effect exceeding that of random 14-protein sets drawn from the plasma proteome (10,000 permutations, $p = 9.99 \times 10^{-5}$). Consistent with this data, from baseline plasma samples, the signature was higher in current compared with former- ($p = 9.3 \times 10^{-10}$) or never-smokers ($p = 5.8 \times 10^{-10}$) in TRACERx (Figure S7D). In addition, a controlled crossover study in healthy subjects[62] that measured 6/14 signature proteins revealed that 2 h of acute diesel exhaust exposure significantly increased plasma levels of 3 myeloid- and fibroblast-associated proteins (MMP12, PLAUR, and TNFSF13B) relative to paired filtered-air control exposure (Table S8).

We next examined how the signature is associated with PM exposure and lung cancer development in predominantly (93.3%) never-smoker participants in TALENT (Figure 1J; STAR Methods). In this cohort, 4/10 (MMP12, PLAUR, TNFSF13B, and LAMP3) measured proteins were associated with high (above median) PM exposure in controls (Figures 3J and S7E). Three of these proteins were induced by acute diesel exposure in the controlled crossover study (Table S8). In future lung cancer cases with above-median PM exposure, 7/10 proteins were elevated compared with controls with low PM exposure (Figures 3J and S7E). The combined signature of these 10 proteins was highest in participants exposed to elevated PM and who developed lung cancer ($p = 0.014$ between high and low PM exposures in future lung cancer cases; Figure S7F). Collectively, these findings suggest that PM and smoking exposure are associated with elevated 14-protein signature levels, consistent with the signature reflecting lung cancer risk from environmental exposures.

The 14-protein signature stratifies patients for lung cancer prevention via anti-IL-1β therapy

Canakinumab reduced lung cancer incidence but was ineffective in established disease,[7–9] indicating a narrow window for interception. As components of the 14-protein signature are elevated prior to diagnosis (Figure 1E) and in response to IL-1β peptide (Figures 3F and S7A), we reasoned that it may identify individuals who benefit from IL-1β blockade. To test this, we performed a retrospective analysis of baseline, 3-, and 12-month serum proteomic data (SomaScan, capturing 10/14 proteins) from participants who had provided separate informed consent at study entry for biomarker sampling in the randomized, double-blind, placebo-controlled, phase 3 CANTOS clinical trial across placebo and canakinumab-treated arms.[63,64]

Across all 4,651 participants, higher levels of the signature were significantly associated with higher lung cancer incidence after controlling for age, smoking status, and body mass index (BMI; median dichotomized signature, hazard ratio [HR] = 2.15, 95% CI: 1.23–3.77, continuous signature HR = 8.82, 95% CI: 2.79–27.82). Individuals in the higher baseline signature group (dichotomized by median signature level) had a greater incidence of lung cancer (62 incident cases/2,326 individuals; 2.67%) than those with a lower baseline signature (17/2,325; 0.73%; Figure 4A). The protein signature exhibited moderate temporal stability in the placebo cohort, with 82% ($N = 1,314$) of participants remaining within the same median-defined category between baseline and 12 months ($\chi^2 = 536$, $p < 0.0016$; Cohen's $\kappa = 0.64$) with a minimal net reclassification index (NRI = −0.02, Figure 4B). The signature was prognostic for lung cancer at each time point within the trial (HR per 1 pooled SD at baseline = 1.609, 95% CI: 1.217–2.127, $p = 0.00085$; at 3 months = 1.554, 95% CI: 1.165–2.071; and at 12 months = 1.690, 95% CI: 1.251–2.282), with no evidence of effect modification between each time point ($p = 0.604$ by Wald test). Together, these data indicate that the protein signature was associated with lung cancer incidence and, importantly, was stable over time.

We next explored whether the signature could identify a group who benefited with reduced cancer incidence after IL-1β blockade. Canakinumab reduced cumulative lung cancer

(G) Linear regression with a 95% CI of the 14-protein and KAC signatures[17] across tissues. Larger points represent mean per condition.
(H) Quantification of immunofluorescence staining for CLDN4$^+$ EGFR-L858R$^+$ cells in lung sections from control and PM-exposed *EGFR*-dox mice, 3 weeks post-oncogene activation ($n = 11–12$ mice per group). Mann-Whitney test.
(I) Quantification of immunofluorescence staining for CLDN4$^+$ cells in PCLS from *EGFR*-dox mice following *ex vivo* PBS control or PM exposure ($n = 8$ mice, paired Wilcoxon test).
(J) Overlap between differentially upregulated genes in different epithelial cell states and LUAD, relative to WT AT2 cells from PBS-exposed control mice (Fisher's test).
(K) Histopathological analyses of *EGFR*-dox mice treated with PBS, 5 or 50 μg PM *in vivo* for 3 weeks and collected at 10 weeks post-induction. WDC, well-differentiated carcinoma; MDC, moderately well-differentiated carcinoma. Two-way ANOVA with Tukey's correction.
(L) Quantification of immunofluorescence staining for CLDN4$^+$ cells in PCLS from *EGFR*-dox mice following *ex vivo* PM exposure and control IgG or anti-IL-1β treatment. Paired Wilcoxon test between PCLS from the same mouse.
(M) Representative human EGFR^{L858R+} immunohistochemistry (IHC) from PM-treated ET mice induced with CMV-Cre, followed by PM treatment exposure for 3 weeks with concomitant IgG control antibody or anti-IL-1β (αIL-1β) treatment and collection at 10 weeks. Quantification of human EGFR-L858R$^+$ lesions per mm^2 of lung tissue, Welch's *t* test, $n = 10$ animals/group.
See also Figures S7 and S9 and Table S9.

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

incidence in the high baseline signature group (3.88% placebo vs. 2.06% canakinumab, odds ratio [OR] 0.52, 95% CI: 0.31–0.86, $p = 0.013$) but not the low-signature group (0.78% vs. 0.72%, OR 0.91, 95% CI: 0.34–2.48, $p = 0.803$; Figure 4C). Accordingly, baseline signature stratification reduced the number needed to treat (NNT) to prevent one additional lung cancer from 1,516 in the low-signature group to 55 (95% CI: 30–343) in the high-signature group (Wald $p = 0.04$; Figure 4D). Consistent with this data, time-to-event analysis showed canakinumab reducing lung cancer hazard in the high-signature group (HR = 0.53, 95% CI: 0.32–0.87) but not the low-signature group (HR = 0.92, 95% CI: 0.34–2.48; continuous signature × treatment interaction $p = 0.19$; Figure 4E).

Given the lack of IL-1β blockade efficacy in established LUAD,[8] we assessed how the signature levels changed in the transition from normal lung to LUAD, reasoning that the temporal window of anti-IL-1β efficacy might coincide with increasing the signature levels. We re-analyzed published bulk RNA-seq data,[65] capturing both malignant epithelial cells and their microenvironment, from 165 surgically resected samples (92% never-smokers, 69% female), including atypical adenomatous hyperplasia (AAH, $N = 25$), adenocarcinoma *in situ* (AIS, $N = 69$), or minimally invasive adenocarcinoma (MIA, $N = 71$), as well as adjacent normal lung tissue ($N = 165$). Data from invasive and non-invasive MIA were analyzed separately.. Expression of genes encoding the 14-protein signature was elevated in AAH compared with normal lung ($p = 1.19 \times 10^{-4}$ by Benjamini-Hochberg-adjusted Dunn's test) and between AIS and AAH ($p = 1.19 \times 10^{-2}$) but remained stable across non-invasive and invasive MIA (Figure 4F). As KACs are thought to represent an intermediate state during malignant transformation,[17] we projected a KAC gene set signature[17] onto the same dataset and observed a positive linear correlation ($R^2 = 0.47$, $p = 3.7 \times 10^{-57}$) between KAC signature expression and the 14-protein signature across all states (Figures 4G and S7G). As components of the 14-protein signature is induced by PM exposure and by IL-1β peptide, its elevation in pre-invasive disease suggests that it captures a tumor-promoting state in which KAC-like malignant trajectories are initiated.

IL-1β blockade limits the malignant potential of KACs

KACs form from multiple lung progenitor cells en route to adenocarcinoma initiation within the alveolar niche (Figure 2). *Kras*-mutant KACs expand in response to IL-1β[23] and drive adenocarcinoma formation.[18,19] Therefore, we explored the effect of PM exposure on KAC evolution and cancer initiation. PM exposure increased the number of Cldn4$^+$ cells, a marker of KAC states, in *EGFR*-mutant lungs *in vivo* and *ex vivo*, an effect not observed in wild-type lungs (Figures 4H, 4I, S7H, and S7I). We next performed snRNA-seq and analyzed 34,459 tdTomato$^+$ cells from PM- and PBS-treated ET and T mice harvested at 3 weeks (small clonal expansions) and 10 weeks (hyperplasia, rare adenoma) after induction with AT2 lineage-restricted adenovirus, focusing on changes in KACs (Figures S8A–S8G; Table S9). PM treatment resulted in upregulation of inflammation and cancer progression-associated pathways[66] at 3 weeks (Myc, tumor necrosis factor α [TNF-α]/nuclear factor κB [NF-κB] signaling, and epithelial-mesenchymal transition) specifically in *EGFR*-mutant KACs (Figure S8G). KAC populations demonstrated overlap with previously reported *Kras*-mutant KAC transcriptional signatures and KAC-like states in bleomycin-induced lung injury[16,17,21] (Figures S8H–S8J). PM-exposed *EGFR*-mutant KACs showed >30% transcriptional overlap (1,298 genes) with late-stage LUAD compared with PBS-treated wild-type AT2 cells (Figure 4J), independent of cell number differences (Figures S9A and S9B). Furthermore, PM exposure accelerated progression to invasive carcinomas *in vivo*, with enhanced formation of carcinomas vs. hyperplasias in the *EGFR*-dox model at 10 weeks compared with PBS-exposed *EGFR*-dox mice (Figures 4K and S9C). These data suggest PM exposure drives *EGFR*-mutant KACs toward transcriptional profiles of adenocarcinoma.

We found that anti-IL-1β treatment reduced Cldn4+ cell numbers in *EGFR*-mutant PCLS exposed to PM (Figure 4L). Anti-IL-1β treatment concurrent with PM exposure *in vivo* reduced the organoid-forming efficiency of *EGFR*-mutant cells (Figure S9D) and limited mutant cell expansions (Figure 4M), consistent with prior results demonstrating reduction of PM-exposed LUADs.[6] In conclusion, we identified a 14-protein signature that reflects a perturbed lung microenvironment, is induced by PM, IL-1β, and *EGFR* oncogenic clones, associates with mutant KACs, and represents a window of opportunity for IL-1β inhibition to prevent incident lung cancer.

DISCUSSION

Here, we identify a 14-protein plasma signature that reflects a tumor-promoting inflammatory alveolar niche, improves risk prediction of future lung cancer, and identifies individuals who may benefit from preventive anti-IL-1β therapy. We show that *EGFR*-driven LUAD can initiate from multiple epithelial lineages that converge on a KAC state within the alveolar niche. Transcriptional elevation of the proteomic signature increases in parallel with a KAC transcriptional signature in pre-invasive disease, consistent with an altered microenvironment that supports KAC emergence. KAC expansion and malignant potential are promoted by PM exposure and inhibited by IL-1β blockade, identifying KACs as a key cellular bottleneck for LUAD prevention.

Transcription of components of the 14-protein signature are inducible by PM or IL-1β exposure in human and murine systems and persists during pre-invasive disease. Unlike multi-cancer early detection assays, which rely on tumor-derived signals and therefore require sufficient lesion burden,[67] the signature is not associated with tumor stage and does not significantly decline following tumor resection in TRACERx. The signature is also elevated in current smokers vs. never-smokers, consistent with the role of tobacco-driven inflammation in carcinogenesis.[68,69] These findings support a model in which the 14-protein signature reflects a lung-specific inflammatory, tumor-promoting state rather than established malignancy and demonstrate the utility of mouse models for mechanistic dissection of risk prediction signatures.[70] Further work is required to define the upstream signaling networks linking mutant epithelial cells and environmental exposures to release these proteins into the circulation,[23,71,72] to determine whether other exposures such as vaping aerosols elevate the signature, and to establish

whether the signature is merely correlative or actively contributes to malignant progression.

The observation that multiple lung epithelial lineages can give rise to *EGFR*-driven adenocarcinoma underscores the role of oncogene-induced plasticity in broadening the pool of tumor-initiating cells[14,15,73] and aligns with human genomic studies proposing basal cells as candidate cells of origin of *EGFR*-driven LUAD.[74] Despite diverse origins, *EGFR*-mutant airway cells migrate into the alveolar niche[75] and subsequently converge on a KAC-like state, suggesting KACs are a key preventive target agnostic to the initiating cell of origin. KACs drive *Kras*-induced malignant progression,[17,18] and by demonstrating that this state is expanded by pollution-induced IL-1β signaling, is sensitive to IL-1β blockade, and is temporally associated with the plasma proteomic signal, our data reinforce the concept that early targeting of the IL-1β axis can intercept lung cancer initiation.[19] Future mechanistic studies are needed to elucidate why anti-IL-1β therapy prevents lung cancer yet shows no efficacy in established disease in the adjuvant or metastatic NSCLC setting.

Retrospective analysis of CANTOS trial participants showed that the 14-protein score stratifies individuals deriving lung cancer preventive benefit from IL-1β inhibition, reducing the NNT to 55, comparable to established cardiovascular prevention strategies.[76] Although exploratory, these data suggest that circulating biomarkers of tumor promotion may enrich prevention trials for individuals most likely to benefit from anti-IL-1β therapy. Together, these data demonstrate how such biomarkers can overcome key barriers to cancer prevention, including the absence of high-penetrance genetic risk markers and the high NNT.[77] Although the signature integrates lung-specific[78] and pleiotropic inflammatory[79] components, replacing the former with other organ-specific plasma signals[80] could enable precision cancer prevention in other cancer types. Collectively, our work highlights the power of integrating machine learning-driven analyses of population-scale proteomics with biological insights from preclinical models and clinical trial data, providing a framework for therapeutic interception and heralding a future for precision cancer prevention.

Limitations of the study

Although respiratory viral exposure is ubiquitous in humans, viral Cre delivery to induce $EGFR^{L858R}$ may introduce inflammatory signals not fully reflective of endogenous oncogene activation. Cldn4-based immunofluorescence likely underestimates KAC heterogeneity, which is more comprehensively defined by single-cell transcriptomics. Among *EGFR*-mutant cells, the KAC state was detected at low frequency (~7.3%–9.8%), and the mechanisms by which such rare populations contribute to systemic proteomic signals remain unclear. The requirement of the KAC state for malignant transformation across different cells of origin and the effects of PM-induced transcriptional changes in KACs need further investigation. UK Biobank analyses were retrospective and lacked baseline CT imaging; however, occult cancer at baseline is unlikely given the median 5-year sampling-to-diagnosis interval and estimated progression times from CT-detectable nodule to asymptomatic stage IA disease.[81] The signature differences between cases and controls in never-smokers in TALENT likely reflect the small sample size of never-smokers in the UK Biobank discovery set, which influenced protein selection, and the shorter time to diagnosis in TALENT compared with the discovery cohort. Ambient $PM_{2.5}$ exposure in human cohorts was estimated by linking residential postcodes to satellite-derived annual averages, which provides an approximation of individual-level exposure and does not account for workplace or lifetime exposures. The proteomic assays provide relative rather than absolute quantification, limiting cross-cohort comparisons. Finally, although the signature-treatment interaction statistical test did not yield $p < 0.05$, the HRs were clinically different (0.53 vs. 0.92). These tests are underpowered in most clinical trials and CANTOS being enriched for cardiovascular disease and not lung cancer.. The signature-group difference in NNT is, nonetheless, hypothesis-generating and motivates signature-stratified enrollment in future prevention trials. Prospective studies incorporating serial sampling and absolute quantification will be required to define actionable thresholds and optimal populations for plasma-guided prevention.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Charles Swanton (charles.swanton@crick.ac.uk).

Materials availability

This study did not generate any new, unique reagents.

Data and code availability

- Murine snRNA-seq, scRNA-seq, bulk RNA-seq, and proteomics data alongside processed TRACERx patient data are deposited on Zenodo at (https://doi.org/10.5281/zenodo.19372114). They are publicly available as of the date of publication.
- UK Biobank data are available to *bona fide* researchers upon application at http://www.ukbiobank.ac.uk/using-the-resource/. The data used in this study were accessed under application number 82693. ARIC data are accessible per the study's data sharing policy (https://sites.cscc.unc.edu/aric/sites/default/files/public/listings/ARIC%20data%20sharing.pdf) and via BioLINCC (controlled access). European Prospective Investigation into Cancer and Nutrition study (EPIC) data access guidelines are available at https://epic.iarc.fr/access/, and EPIC-Norfolk data can be requested by *bona fide* researchers for specified scientific purposes via the study website (https://www.mrc-epid.cam.ac.uk/research/studies/epic-norfolk/). CKB data are available to the international scientific community via application at https://www.ckbiobank.org/data-access. TALENT study data can be requested by contacting Professor P.C. Yang (pcyang@ntu.edu.tw). CANTOS individual participant data cannot be publicly deposited due to informed consent and regulatory restrictions. Qualified researchers may apply for access to anonymized patient-level data through Novartis via an independent scientific review process. UKCTOCS data cannot be deposited in a public repository given the terms of consent. Access requests should be directed to Professor Usha Menon, UCL (u.menon@ucl.ac.uk). TRACERx data access is controlled by the TRACERx data access committee. Details on how to apply for access are available on the linked page.
- All original code has been deposited at Zenodo at (https://doi.org/10.5281/zenodo.19372114) and is publicly available as of the date of publication.
- Any additional information required to re-analyze the data reported in this paper is available from the lead contact upon request.

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

ACKNOWLEDGMENTS

The authors thank the staff of the facilities at The Francis Crick Institute (Flow Cytometry, Experimental Histopathology, Advanced Light Microscopy, and the Genomics Science Technology Platforms) as well as members of the TRACERx consortium for their contributions. TRACERx is funded by Cancer Research UK (CRUK; C11496/A17786) and is coordinated by CRUK and the UCL Cancer Trials Centre. The authors thank the staff and participants of the ARIC study and UKCTOCS trial for their important contributions. UKCTOCS was funded by the Medical Research Council (G9901012 and G0801228), CRUK (C1479/A2884), the National Institute for Health Research (16/46/01), and The Eve Appeal. ARIC has been funded in whole or in part with federal funds from the National Heart, Lung, and Blood Institute; the National Institutes of Health; and the Department of Health and Human Services (under contract nos. 75N92022D00001, 75N92022D00002, 75N92022D00003, 75N92022D00004, and 75N92022D00005). Studies on cancer in ARIC are supported by the National Cancer Institute (U01 CA164975). SomaLogic Inc. conducted the SomaScan assays in exchange for use of ARIC data. ARIC work was also supported in part by NIH/NHLBI grant R01 HL134320. Cancer data were provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry data are also supported by Cooperative Agreement NU58DP007114, funded by the Centers for Disease Control and Prevention. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the Centers for Disease Control and Prevention, the National Institutes of Health, or the Department of Health and Human Services. We thank Emma Rawlins and her lab for providing fetal-lung-derived alveolar organoids made available by the Joint MRC/Wellcome Trust (grant no. MR/X008304/1 and 226202/Z/22/Z) Human Developmental Biology Resource (http://hdbr.org). We acknowledge the extensive collaboration and dedication required to keep the Air Pollution Exposure Lab (APEL) operational to generate the high-quality data included in this paper. The coordination of EPIC-Europe is financially supported by the International Agency for Research on Cancer and by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London. The national cohorts are supported by Associazione Italiana per la Ricerca Sul Cancro-AIRC-Italy, the Italian Ministry of Health, the Italian Ministry of University and Research, and Compagnia di San Paolo (Italy); the Dutch Ministry of Public Health, Welfare and Sports, the Netherlands Organisation for Health Research and Development, and the World Cancer Research Fund; the Instituto de Salud Carlos III (ISCIII), the Regional Governments of Andalucía, Asturias, Basque Country, Murcia, and Navarra, and the Catalan Institute of Oncology – ICO (Spain); and Cancer Research UK (C864/A14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford) and the Medical Research Council (MR/N003284/1, MC-UU_12015/1, and MC_UU_00006/1 to EPIC-Norfolk; MR/Y013662/1 to EPIC-Oxford) (United Kingdom). SomaScan data were generated under the Master Research Agreement, 14th December 2021, between Imperial College London and SomaLogic Inc. SomaLogic was not involved in analyzing or interpreting the data or in writing or submitting the manuscript for publication. C.S. is a Royal Society Napier Research Professor (RSRP\R\210001). C.S., C.S.I., and A.W. are supported by The Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2041, CC2085 for C.S.I. and A.W.), the UK Medical Research Council (CC2041, CC2085 for C.S.I. and A.W.), and the Wellcome Trust (CC2041, CC2085 for C.S.I. and A.W.). For the purpose of open access, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission. C.S. is funded by Cancer Research UK (TRACERx [C11496/A17786], PEACE [C416/A21999], and CRUK Cancer Immunotherapy Catalyst Network); the Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025); the Rosetrees Trust, Butterfield, and Stoneygate Trusts; the NovoNordisk Foundation (ID16584); the Royal Society Professorship Enhancement Award (RP/EA/180007 and RF\ERE\231118); the National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre; the Cancer Research UK-University College London Centre; the Experimental Cancer Medicine Centre; the Breast Cancer Research Foundation (US) (BCRF-23-157); the Cancer Research UK Early Detection and Diagnosis Primer Award (grant EDDPMA-Nov21/100034); and the Mark Foundation for Cancer Research Aspire Award (grant 21-029-ASP) and ASPIRE Phase II award (grant 23-034-ASP). C.S., E.G., and W.H. are funded by an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 835297) and START, a UKRI Frontier Research grant (reference EP/Z534298/1). M.Z. was supported by the EMBO Postdoctoral Fellowship (ALTF 831-2023). T.P. is supported by a UCL UKRI Centre for Doctoral Training in AI-enabled Healthcare studentship (EP/S021612/1), a Francis Crick Institute Idea to Innovation (i2i) grant, the Ruth Strauss Foundation, and a Cancer Research Horizons Translational Fund award. M.A. was supported by the City of London Center Clinical Academic Training Program (year 3, SEBSTF-2021\100007). L.Y.L. is supported by the Canadian Institutes of Health Research Banting Postdoctoral Fellowship. A.J.G. was supported by Einstein's CMBG T32 training grant (5T32GM145438) and is supported by an NCI F31 fellowship (1F31CA295001). O.B. acknowledges support from Barts Charity (G-001522). O.B., T.C.-J., U.M., S.A., A.Z., R.G., Z.L., and H.J.W. acknowledge support from UK/EPSRC joint award EDDCPJT/100022. A.Z. acknowledges the Basic Research Program at the National Research University Higher School of Economics (HSE University) and its HPC facilities, Moscow, Russia. U.M. acknowledges support from MRC CTU at UCL core funding (MC_UU_00004/01). A.G. was supported by the MSCA/UKRIPostdoctoral Fellowship (EP/Y031091/1). C.G. is a CRUK Clinical Research Training Fellow, supported by the CRUK City of London Centre Award (SEBCATP-2024/100003). T. Karasaki is supported by the Japan Society for the Promotion of Science (JSPS) overseas research fellowships program (202060447). S.F.'s work was financially supported by the Brain Cancer Centre (Carrie's Beanies 4 Brain Cancer). L.M.L. is supported by the Lung Cancer Research Foundation, A Breath of Hope Foundation, The American Association for Cancer Research, and the V Foundation. L.M.L. is a Jane A. and Myles P. Dempsey Cancer Research Scholar and holds the Rubenstein Family Endowed Early Career Professorship in Environmental Determinants and Disease. N. McGranahan receives funding from Cancer Research UK (CRUK) (DRCPFA-Nov23/100003), the Wellcome Trust, the Royal Society (211179/Z/18/Z), CRUK LCCE, Rosetrees, and the NIHR BRC at University College London Hospitals. W.H. is funded by a University of Manchester's Vice Chancellor and Cancer Research UK Manchester Institute Fellowship (C5759/A27412), a CRUK EDD (EDDPMA-Nov21\100034), the Mark Foundation for Cancer Research Aspire Award (grant 21-029-ASP), and the ASPIRE Phase II award (grant 23-034-ASP). C.E.W. was supported by the European Respiratory Society and the European Union's H2020 research and innovation program under Marie Sklodowska-Curie (grant 847462), CRUK EDD (EDDPMA-Nov21\100034), The Mark Foundation for Cancer Research ASPIRE Phase II award (grant 23-034-ASP), and a CSL Centenary Fellowship 2025-2029. L.M. is a national Mah Jongg League Fellow of the Damon Runyon Cancer Research Foundation (DRG-2560-25). K.H.C. is supported by a Wellcome Career Development Award (315647/Z/24/Z). J.Z. is funded by the Cancer Prevention and Research Institute of Texas (CPRIT) Clinical Investigator award RP240441 and the MD Anderson Lung Cancer Interception Program. A.J.-S. is supported by an AACR-CRUK Transatlantic Fellowship (AACR1766).

AUTHOR CONTRIBUTIONS

Conceptualization, T.P., M.Z., M.M.L., M.A., W.H., C.E.W., E.G., and C.S.; formal analysis, T.P., M.Z., M.M.L., M.A., C.E.W., W.H., A.-M.L., U.B., T. Klockner, O.B., K.H., D.C.M., K.H.C., V.A.B., N.W., J.W., E.P., S.W.N., C.G., and A.A.; funding acquisition, T.P., M.Z., M.A., O.B., U.M., K.L., N.R.N., N. McGranahan, E.G., W.H., C.E.W., and C.S.; methodology, T.P., M.Z., M.M.L., M. Mugabo, M.A., C.E.W., W.H., L.Y.L., A.-M.L., U.B., K.H., E.P., J.W., R.M., K.L., N.R.N., N. McGranahan, E.G., D.C.M., H.Z., L.M., and A.C.; resources, T.P., M.M.L., A.-M.L., T. Klockner, M. Merad, O.B., C.S.I., A. Guenoun, J.W., N.W., A.L.M., S.W., R.C.H.V., P.M.K., H.Z., T. Karasaki, M.G., N.C., E.A.P., K.S.-B., W.H., M.Z., M.J.-H., C.E.W., C.S., A. Grenov, T.S., A.S.-B., S.L.P., A.H., P.-C.Y., M.J., S.A., and U.M.; supervision, C.E.W., W.H., L.Y.L., A.J.G., C.S., N.R.N., N. McGranahan, E.G., A.W., M.Z., P.-C.Y., K.S.-B., and S.F.; writing – original draft, T.P., M.Z., M.A., M.M.L., C.E.W.,

W.H., C.S., and E.G.; and writing – review and editing, T.P., M.Z., M.A., E.G., L.Y.L., A.J.G., C.G., E.A.P., K.S.-B., M.M.L., L.M.L., N. McGranahan, C.E.W., W.H., E.S., U.M., and C.S.

DECLARATION OF INTERESTS

C.S. acknowledges grant support from AstraZeneca, Boehringer Ingelheim, Bristol Myers Squibb, Pfizer, Invitae (previously Archer Dx Inc.—collaboration in minimal residual disease sequencing technologies), Ono Pharmaceutical, and Personalis. He is also Co-Chief Investigator of the NHS Galleri trial funded by GRAIL and a paid member of GRAIL's Scientific Advisory Board. He was Chief Investigator for the AZ MeRmaiD 1 and 2 clinical trials and the Steering Committee Chair. C.S. has been a non-executive board member for Novartis since March 2026. He is also a paid board member for Bicycle Therapeutics and is Chair of the Clinical Advisory Group. He receives consultant fees from Genentech, Medicxi, China Innovation Centre of Roche (CICoR) (formerly Roche Innovation Centre - Shanghai), Relay Therapeutics (SAB member), Saga Diagnostics (SAB member), and Sarah Cannon Research Institute. He previously received consultant fees from Achilles Therapeutics. C.S. has received honoraria from Amgen, AstraZeneca, Bristol Myers Squibb, GlaxoSmithKline, Illumina, MSD, Novartis, and Pfizer. C.S. has equity in Bicycle Therapeutics, Relay Therapeutics, and Novartis. He has stock options in Relay Therapeutics, Saga Diagnostics, and Bicycle Therapeutics. He had previously held stock and was co-founder of Achilles Therapeutics. C.S. declares a patent application for methods to detect lung cancer (PCT/US2017/028013), targeting neoantigens (PCT/EP2016/059401), identifying patient response to immune checkpoint blockade (PCT/EP2016/071471), methods for lung cancer detection (US20190106751A1), identifying patients who respond to cancer treatment (PCT/GB2018/051912), determining HLA LOH (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221), methods and systems for tumor monitoring (PCT/EP2022/077987), and analysis of HLA alleles transcriptional deregulation (PCT/EP2023/059039). C.S. is an inventor on a European patent application (PCT/GB2017/053289) relating to assay technology to detect tumor recurrence. This patent has been licensed to a commercial entity, and under their terms of employment, C.S. is due a revenue share of any revenue generated from such license(s). T.P., M.A., O.B., N.R.N., W.H., M.M.L., C.E.W., M.Z., L.Y.L., and C.S. are named as inventors on PCT/EP2025/086701 relating to use of plasma proteomics for risk prediction of lung cancer. T.P. and M.A. have undertaken consultancy work for FutureHouse and Edison Scientific, unrelated to this project. T.P., M.A., N.R.N., and C.S. are listed as inventors on a further patent application (GB) that has been filed related to methods described in this manuscript, unpublished, and remains within the priority year. M.Z. is employed by Baseimmune Ltd. and holds share options in the company; this employment is unrelated to this study. U.M. and S.A. report personal consulting fees from Mercy BioAnalytics Ltd. and research support grants paid to the institution from Intelligent Lab on Fiber, RNA Guardian, MercyBio Analytics, and CleoDx related to early detection of cancer, especially ovarian cancer. U.M. declares membership of the Research Advisory Panel, Yorkshire Cancer Research (UK), and honorarium for membership of Tina's Wish Scientific Advisory Board (USA). U.M. holds patent number EP10178345.4 for Breast Cancer Diagnostics. T.C.-J. is a member of the Advisory Board of Early Health and holds a patent EP3254111A1 (Biomarkers for pancreatic cancer detection). Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization. M.J.-H. has received funding from CRUK, the NIH National Cancer Institute, the IASLC International Lung Cancer Foundation, the Lung Cancer Research Foundation, the Rosetrees Trust, UKI NETs, and the NIHR. M.J.-H. has consulted for Astex Pharmaceuticals, Pfizer, and Achilles Therapeutics; is a member of the Achilles Therapeutics Scientific Advisory Board and Steering Committee; and has received speaker honoraria from Pfizer, Astex Pharmaceuticals, Oslo Cancer Cluster, Bristol Myers Squibb, Genentech, and GenesisCare. M.J.-H. is listed as a co-inventor on a European patent application relating to methods to detect lung cancer (PCT/US2017/028013), and this patent has been licensed to commercial entities, and, under terms of employment, M.J.-H. is due a share of any revenue generated from such license(s) and is also listed as a co-inventor on the GB priority patent application (GB2400424.4) with the title Treatment and Prevention of Lung Cancer. K.L. has a patent on indel burden and CPI response pending, and, outside of the submitted work, speaker fees from Roche tissue diagnostics, research funding from the CRUK TDL/Ono/LifeArc alliance and Genesis Therapeutics, and consulting roles with Monopteros Therapeutics, Kynos Therapeutics, Saga Diagnostics, and Tempus and is currently a full-time employee at Isomorphic Labs, all outside of the submitted work. N. McGranahan holds patents related to determining HLA LOH (PCT/GB2018/052004), determination of B cell fraction in mixed samples (PCT/EP2024/062999), determination of lymphocyte abundance in mixed samples (PCT/EP2022/070694), identifying responders to cancer treatment (PCT/GB2018/051912), targeting neoantigens (PCT/EP2016/059401), identifying patient response to immune checkpoint blockade (PCT/EP2016/071471), and predicting survival rates of patients with cancer (PCT/GB2020/050221) and has a patent pending in determining HLA disruption (PCT/EP2023/059039). J.Z. reports research funding from Merck, Johnson and Johnson, Fortvita, Novartis, Summit, and Henlius and consultant fees from ASP, Delcath, Johnson and Johnson, AstraZeneca, Innovent, Varian, and Catalyst outside the submitted work. Dr. Mattias Johansson reported being a named inventor on a provisional patent application for the INTEGRAL-Risk model but will not receive any royalties or other compensation related to this potential patent. Dr. Hana Zahed reported being a named inventor on a provisional patent application for the INTEGRAL-Risk model but will not receive any royalties or other compensation related to this potential patent.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Mouse models
  - Mouse lung organoids and cell lines
  - Murine lung organoid dataset
  - UK Biobank
  - External validation datasets
  - TRACERx study
  - CANTOS study
  - TALENT study
  - Air pollution exposure study
  - Bulk human pre-invasive RNA-seq
  - Human foetal lung-derived alveolar organoids
- METHOD DETAILS
  - Animal procedures
  - MicroCT imaging
  - Histopathology and immuno-staining
  - Flow cytometry
  - Single-nuclei RNA sequencing
  - snRNA-seq preprocessing
  - snRNA-seq QC, clustering and annotation
  - Single-cell RNA sequencing (scRNA-seq)
  - scRNA-seq preprocessing
  - scRNA-seq QC, clustering and annotation
  - Analysis of murine lung organoid scRNA-seq
  - PCLS generation for live cell imaging
  - *Ex vivo* PM challenge of PCLS
  - Bulk RNA-seq of murine tumor tissue
  - Bulk RNA-seq analysis of pre-invasive dataset
  - Machine learning model development
  - Model Validation
  - Protein measurement in validation cohorts
  - UKCTOCS
  - LC3 Consortium
  - deCODE genetics + Icelandic Cancer Project
  - EPIC

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2026.05.005.

Received: July 4, 2025
Revised: January 22, 2026
Accepted: May 6, 2026

REFERENCES

1. Lam, S., Bai, C., Baldwin, D.R., Chen, Y., Connolly, C., de Koning, H., Heuvelmans, M.A., Hu, P., Kazerooni, E.A., Lancaster, H.L., et al. (2024). Current and Future Perspectives on Computed Tomography Screening for Lung Cancer: A Roadmap From 2023 to 2027 From the International Association for the Study of Lung Cancer. J. Thorac. Oncol. *19*, 36–51. https://doi.org/10.1016/j.jtho.2023.07.019.
2. Krilaviciute, A., and Brenner, H. (2021). Low positive predictive value of computed tomography screening for lung cancer irrespective of commonly employed definitions of target population. Int. J. Cancer *149*, 58–65. https://doi.org/10.1002/ijc.33522.
3. LoPiccolo, J., Gusev, A., Christiani, D.C., and Jänne, P.A. (2024). Lung cancer in patients who have never smoked - an emerging disease. Nat. Rev. Clin. Oncol. *21*, 121–146. https://doi.org/10.1038/s41571-023-00844-0.
4. Yoshida, K., Gowers, K.H.C., Lee-Six, H., Chandrasekharan, D.P., Coorens, T., Maughan, E.F., Beal, K., Menzies, A., Millar, F.R., Anderson, E., et al. (2020). Tobacco smoking and somatic mutations in human bronchial epithelium. Nature *578*, 266–272. https://doi.org/10.1038/s41586-020-1961-1.
5. Kakiuchi, N., and Ogawa, S. (2021). Clonal expansion in non-cancer tissues. Nat. Rev. Cancer *21*, 239–256. https://doi.org/10.1038/s41568-021-00335-3.
6. Hill, W., Lim, E.L., Weeden, C.E., Lee, C., Augustine, M., Chen, K., Kuan, F.-C., Marongiu, F., Evans, E.J., Moore, D.A., et al. (2023). Lung adenocarcinoma promotion by air pollutants. Nature *616*, 159–167. https://doi.org/10.1038/s41586-023-05874-3.
7. Ridker, P.M., MacFadyen, J.G., Thuren, T., Everett, B.M., Libby, P., Glynn, R.J., Ridker, P., Lorenzatti, A., Krum, H., Varigos, J., et al. (2017). Effect of interleukin-1β inhibition with canakinumab on incident lung cancer in patients with atherosclerosis: exploratory results from a randomised, double-blind, placebo-controlled trial. Lancet *390*, 1833–1842. https://doi.org/10.1016/S0140-6736(17)32247-X.
8. Garon, E.B., Lu, S., Goto, Y., De Marchi, P., Paz-Ares, L., Spigel, D.R., Thomas, M., Yang, J.C.-H., Ardizzoni, A., Barlesi, F., et al. (2024). Canakinumab as Adjuvant Therapy in Patients With Completely Resected Non–Small-Cell Lung Cancer: Results From the CANOPY-A Double-Blind, Randomized Clinical Trial. J. Clin. Oncol. *42*, 180–191. https://doi.org/10.1200/JCO.23.00910.
9. Tan, D.S.W., Felip, E., de Castro, G., Solomon, B.J., Greystoke, A., Cho, B.C., Cobo, M., Kim, T.M., Ganguly, S., Carcereny, E., et al. (2024). Canakinumab Versus Placebo in Combination With First-Line Pembrolizumab Plus Chemotherapy for Advanced Non–Small-Cell Lung Cancer: Results From the CANOPY-1 Trial. J. Clin. Oncol. *42*, 192–204. https://doi.org/10.1200/JCO.23.00980.
10. Nikolić, M.Z., and Hogan, B.L.M., eds. (2021). Lung Stem Cells in Development, Health and Disease, First Edition (European Respiratory Society). https://doi.org/10.1183/2312508X.erm9121.
11. Ferone, G., Lee, M.C., Sage, J., and Berns, A. (2020). Cells of origin of lung cancers: lessons from mouse studies. Genes Dev. *34*, 1017–1032. https://doi.org/10.1101/gad.338228.120.
12. Xu, X., Rock, J.R., Lu, Y., Futtner, C., Schwab, B., Guinney, J., Hogan, B.L.M., and Onaitis, M.W. (2012). Evidence for type II cells as cells of origin of K-Ras-induced distal lung adenocarcinoma. Proc. Natl. Acad. Sci. USA *109*, 4910–4915. https://doi.org/10.1073/pnas.1112499109.
13. Sutherland, K.D., Song, J.-Y., Kwon, M.C., Proost, N., Zevenhoven, J., and Berns, A. (2014). Multiple cells-of-origin of mutant K-Ras-induced mouse lung adenocarcinoma. Proc. Natl. Acad. Sci. USA *111*, 4952–4957. https://doi.org/10.1073/pnas.1319963111.
14. Spella, M., Lilis, I., Pepe, M.A., Chen, Y., Armaka, M., Lamort, A.-S., Zazara, D.E., Roumelioti, F., Vreka, M., Kanellakis, N.I., et al. (2019). Club cells form lung adenocarcinomas and maintain the alveoli of adult mice. eLife *8*, e45571. https://doi.org/10.7554/eLife.45571.
15. Juul, N.H., Yoon, J.-K., Martinez, M.C., Rishi, N., Kazadaeva, Y.I., Morri, M., Neff, N.F., Trope, W.L., Shrager, J.B., Sinha, R., et al. (2023). KRAS(G12D) drives lepidic adenocarcinoma through stem-cell reprogramming. Nature *619*, 860–867. https://doi.org/10.1038/s41586-023-06324-w.
16. Marjanovic, N.D., Hofree, M., Chan, J.E., Canner, D., Wu, K., Trakala, M., Hartmann, G.G., Smith, O.C., Kim, J.Y., Evans, K.V., et al. (2020). Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. Cancer Cell *38*, 229–246.e13. https://doi.org/10.1016/j.ccell.2020.06.012.
17. Han, G., Sinjab, A., Rahal, Z., Lynch, A.M., Treekitkarnmongkol, W., Liu, Y., Serrano, A.G., Feng, J., Liang, K., Khan, K., et al. (2024). An atlas of epithelial cell states and plasticity in lung adenocarcinoma. Nature *627*, 656–663. https://doi.org/10.1038/s41586-024-07113-9.
18. Chan, J.E., Pan, C.-H., Rub, J., Guzman, G., Krause, K., Brown, E., Zhang, Z., Styers, H., Hartmann, G., Li, Z., et al. (2026). Critical role for a high-plasticity cell state in lung cancer. Nature *651*, 231–241. https://doi.org/10.1038/s41586-025-09985-x.
19. Peng, F., Sinjab, A., Dai, Y., Treekitkarnmongkol, W., Yang, S., Gomez Bolanos, L.I.G., Zhou, T., Chen, M., Serrano, A.G., Krishna, A., et al. (2026). Multimodal spatial-omics reveal co-evolution of alveolar progenitors and proinflammatory niches in progression of lung precursor lesions. Cancer Cell *44*, 321–339.e13. https://doi.org/10.1016/j.ccell.2025.10.004.
20. Strunz, M., Simon, L.M., Ansari, M., Kathiriya, J.J., Angelidis, I., Mayr, C.H., Tsidiridis, G., Lange, M., Mattner, L.F., Yee, M., et al. (2020). Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. Nat. Commun. *11*, 3559. https://doi.org/10.1038/s41467-020-17358-3.
21. Choi, J., Park, J.-E., Tsagkogeorga, G., Yanagita, M., Koo, B.-K., Han, N., and Lee, J.-H. (2020). Inflammatory Signals Induce AT2 Cell-Derived Damage-Associated Transient Progenitors that Mediate Alveolar Regeneration. Cell Stem Cell *27*, 366–382.e7. https://doi.org/10.1016/j.stem.2020.06.020.
22. Kobayashi, Y., Tata, A., Konkimalla, A., Katsura, H., Lee, R.F., Ou, J., Banovich, N.E., Kropski, J.A., and Tata, P.R. (2020). Persistence of a regeneration-associated, transitional alveolar epithelial cell state in pulmonary fibrosis. Nat. Cell Biol. *22*, 934–946. https://doi.org/10.1038/s41556-020-0542-8.
23. England, F.J., Bordeu, I., Ng, M.-E., Bang, J., Kim, B., Choi, J., Cardoso, E.C., Koo, B.-K., Simons, B.D., and Lee, J.-H. (2025). Sustained NF-κB

activation allows mutant alveolar stem cells to co-opt a regeneration program for tumor initiation. Cell Stem Cell *32*, 375–390.e9. https://doi.org/10.1016/j.stem.2025.01.011.

24. Narasimhan, H., Cheon, I.S., Qian, W., Hu, S.S., Parimon, T., Li, C., Goplen, N., Wu, Y., Wei, X., Son, Y.M., et al. (2024). An aberrant immune-epithelial progenitor niche drives viral lung sequelae. Nature *634*, 961–969. https://doi.org/10.1038/s41586-024-07926-8.

25. Lang, N.J., Gote-Schniering, J., Porras-Gonzalez, D., Yang, L., De Sadeleer, L.J., Jentzsch, R.C., Shitov, V.A., Zhou, S., Ansari, M., Agami, A., et al. (2023). Ex vivo tissue perturbations coupled to single-cell RNA-seq reveal multilineage cell circuit dynamics in human lung fibrogenesis. Sci. Transl. Med. *15*, eadh0908. https://doi.org/10.1126/scitranslmed.adh0908.

26. Adams, T.S., Schupp, J.C., Poli, S., Ayaub, E.A., Neumark, N., Ahangari, F., Chu, S.G., Raby, B.A., Deluliis, G., Januszyk, M., et al. (2020). Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci. Adv. *6*, eaba1983. https://doi.org/10.1126/sciadv.aba1983.

27. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.-H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. Nature *622*, 329–338. https://doi.org/10.1038/s41586-023-06592-6.

28. Pisabarro, M.T., Leung, B., Kwong, M., Corpuz, R., Frantz, G.D., Chiang, N., Vandlen, R., Diehl, L.J., Skelton, N., Kim, H.S., et al. (2006). Cutting edge: novel human dendritic cell- and monocyte-attracting chemokine-like protein identified by fold recognition methods. J. Immunol. *176*, 2069–2073. https://doi.org/10.4049/jimmunol.176.4.2069.

29. Khan, T., Kryza, T., Lyons, N.J., He, Y., and Hooper, J.D. (2021). The CDCP1 Signaling Hub: A Target for Cancer Detection and Therapeutic Intervention. Cancer Res. *81*, 2259–2269. https://doi.org/10.1158/0008-5472.CAN-20-2978.

30. Lun, Y., Borjini, N., Miura, N.N., Ohno, N., Singer, N.G., and Lin, F. (2021). CDCP1 on Dendritic Cells Contributes to the Development of a Model of Kawasaki Disease. J. Immunol. *206*, 2819–2827. https://doi.org/10.4049/jimmunol.2001406.

31. Zhang, Y., Jiang, M., Nouraie, M., Roth, M.G., Tabib, T., Winters, S., Chen, X., Sembrat, J., Chu, Y., Cardenes, N., et al. (2019). GDF15 is an epithelial-derived biomarker of idiopathic pulmonary fibrosis. Am. J. Physiol., Lung Cell. Mol. Physiol. *317*, L510–L521. https://doi.org/10.1152/ajplung.00062.2019.

32. Planté-Bordeneuve, T., Bertrand, Y., Lecocq, M., Hoton, D., Fillée, C., Lacroix, V., Rondelet, B., Wuyts, W., Bouzin, C., Pilette, C., et al. (2024). The IgA-pIgR System Is Dysregulated in Idiopathic Pulmonary Fibrosis. Am. J. Respir. Crit. Care Med. *210*, 841–844. https://doi.org/10.1164/rccm.202401-0043LE.

33. Nascimento, M., Huot-Marchand, S., Gombault, A., Panek, C., Bourinet, M., Fanny, M., Savigny, F., Schneider, P., Le Bert, M., Ryffel, B., et al. (2020). B-Cell Activating Factor Secreted by Neutrophils Is a Critical Player in Lung Inflammation to Cigarette Smoke Exposure. Front. Immunol. *11*, 1622. https://doi.org/10.3389/fimmu.2020.01622.

34. Smith, H.W., and Marshall, C.J. (2010). Regulation of cell signalling by uPAR. Nat. Rev. Mol. Cell Biol. *11*, 23–36. https://doi.org/10.1038/nrm2821.

35. Shapiro, S.D., Kobayashi, D.K., and Ley, T.J. (1993). Cloning and characterization of a unique elastolytic metalloproteinase produced by human alveolar macrophages. J. Biol. Chem. *268*, 23824–23829. https://doi.org/10.1016/S0021-9258(20)80459-1.

36. Beauchemin, N., and Arabzadeh, A. (2013). Carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) in cancer progression and metastasis. Cancer Metastasis Rev. *32*, 643–671. https://doi.org/10.1007/s10555-013-9444-6.

37. Bingle, L., Cross, S.S., High, A.S., Wallace, W.A., Rassl, D., Yuan, G., Hellstrom, I., Campos, M.A., and Bingle, C.D. (2006). WFDC2 (HE4): a potential role in the innate immunity of the oral cavity and respiratory tract and the development of adenocarcinomas of the lung. Respir. Res. *7*, 61. https://doi.org/10.1186/1465-9921-7-61.

38. Chen, Y., Dou, R., Hong, M.J., Xu, H., Vykoukal, J., León-Letelier, R.A., Cai, Y., Park, S., Irajizad, E., Hsiao, F.C., et al. (2025). Lung adenocarcinoma surfaceome remodeling with EGFR inhibitors uncovers placental alkaline phosphatase as a target for combination therapy. Cell Rep. Med. *6*, 102513. https://doi.org/10.1016/j.xcrm.2025.102513.

39. Verghese, G.M., Tong, Z.Y., Bhagwandin, V., and Caughey, G.H. (2004). Mouse prostasin gene structure, promoter analysis, and restricted expression in lung and kidney. Am. J. Respir. Cell Mol. Biol. *30*, 519–529. https://doi.org/10.1165/rcmb.2003-0251OC.

40. Lunding, L.P., Krause, D., Stichtenoth, G., Stamme, C., Lauterbach, N., Hegermann, J., Ochs, M., Schuster, B., Sedlacek, R., Saftig, P., et al. (2021). LAMP3 deficiency affects surfactant homeostasis in mice. PLoS Genet. *17*, e1009619. https://doi.org/10.1371/journal.pgen.1009619.

41. Han, S., and Mallampalli, R.K. (2015). The Role of Surfactant in Lung Disease and Host Defense against Pulmonary Infections. Ann. Am. Thorac. Soc. *12*, 765–774. https://doi.org/10.1513/AnnalsATS.201411-507FR.

42. Menon, U., Gentry-Maharaj, A., Burnell, M., Singh, N., Ryan, A., Karpinskyj, C., Carlino, G., Taylor, J., Massingham, S.K., Raikou, M., et al. (2021). Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet *397*, 2182–2193. https://doi.org/10.1016/S0140-6736(21)00731-5.

43. Albanes, D., Alcala, K., Alcala, N., Amos, C.I., Arslan, A.A., Bassett, J.K., Brennan, P., Cai, Q., Chen, C., Feng, X., et al. (2023). The blood proteome of imminent lung cancer diagnosis. Nat. Commun. *14*, 3042. https://doi.org/10.1038/s41467-023-37979-8.

44. Eldjarn, G.H., Ferkingstad, E., Lund, S.H., Helgason, H., Magnusson, O.T., Gunnarsdottir, K., Olafsdottir, T.A., Halldorsson, B.V., Olason, P.I., Zink, F., et al. (2023). Large-scale plasma proteomics comparisons through genetics and disease associations. Nature *622*, 348–358. https://doi.org/10.1038/s41586-023-06563-x.

45. Carrasco-Zanini, J., Pietzner, M., Koprulu, M., Wheeler, E., Kerrison, N.D., Wareham, N.J., and Langenberg, C. (2024). Proteomic prediction of diverse incident diseases: a machine learning-guided biomarker discovery study using data from a prospective cohort study. Lancet Digit. Health *6*, e470–e479. https://doi.org/10.1016/S2589-7500(24)00087-6.

46. Lind, L., Mazidi, M., Clarke, R., Bennett, D.A., and Zheng, R. (2024). Measured and genetically predicted protein levels and cardiovascular diseases in UK Biobank and China Kadoorie Biobank. Nat CardioVasc Res. *3*, 1189–1198. https://doi.org/10.1038/s44161-024-00545-6.

47. Shelbaya, K., Arthur, V., Yang, Y., Dorbala, P., Buckley, L., Claggett, B., Skali, H., Dufresne, L., Yang, T.-Y., Engert, J.C., et al. (2024). Large-Scale Proteomics Identifies Novel Biomarkers and Circulating Risk Factors for Aortic Stenosis. J. Am. Coll. Cardiol. *83*, 577–591. https://doi.org/10.1016/j.jacc.2023.11.021.

48. Ferkingstad, E., Sulem, P., Atlason, B.A., Sveinbjornsson, G., Magnusson, M.I., Styrmisdottir, E.L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B.V., et al. (2021). Large-scale integration of the plasma proteome with genetics and disease. Nat. Genet. *53*, 1712–1721. https://doi.org/10.1038/s41588-021-00978-w.

49. Sikkema, L., Ramírez-Suástegui, C., Strobl, D.C., Gillett, T.E., Zappia, L., Madissoon, E., Markov, N.S., Zaragosi, L.-E., Ji, Y., Ansari, M., et al. (2023). An integrated cell atlas of the lung in health and disease. Nat. Med. *29*, 1563–1577. https://doi.org/10.1038/s41591-023-02327-2.

50. Aguet, F., Anand, S., Ardlie, K.G., Gabriel, S., Getz, G.A., Graubert, A., Hadley, K., Handsaker, R.E., Huang, K.H., Kashin, S., et al. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

51. Feng, X., Goodley, P., Alcala, K., Guida, F., Kaaks, R., Vermeulen, R., Downward, G.S., Bonet, C., Colorado-Yohar, S.M., Albanes, D., et al. (2024). Evaluation of risk prediction models to select lung cancer screening participants in Europe: a prospective cohort consortium analysis. Lancet Digit. Health *6*, e614–e624. https://doi.org/10.1016/S2589-7500(24)00123-7.

52. Field, J.K., Vulkan, D., Davies, M.P.A., Duffy, S.W., and Gabe, R. (2021). Liverpool Lung Project lung cancer risk stratification model: calibration and prospective validation. Thorax *76*, 161–168. https://doi.org/10.1136/thoraxjnl-2020-215158.

53. Al-Sawaf, O., Weiss, J., Skrzypski, M., Lam, J.M., Karasaki, T., Zambrana, F., Kidd, A.C., Frankell, A.M., Watkins, T.B.K., Martínez-Ruiz, C., et al. (2023). Body composition and lung cancer-associated cachexia in TRACERx. Nat. Med. *29*, 846–858. https://doi.org/10.1038/s41591-023-02232-8.

54. Chang, G.-C., Chiu, C.-H., Yu, C.-J., Chang, Y.-C., Chang, Y.-H., Hsu, K.-H., Wu, Y.-C., Chen, C.-Y., Hsu, H.-H., Wu, M.-T., et al. (2024). Low-dose CT screening among never-smokers with or without a family history of lung cancer in Taiwan: a prospective cohort study. Lancet Respir. Med. *12*, 141–152. https://doi.org/10.1016/S2213-2600(23)00338-7.

55. Sutherland, K.D., Proost, N., Brouns, I., Adriaensen, D., Song, J.-Y., and Berns, A. (2011). Cell of origin of small cell lung cancer: inactivation of Trp53 and Rb1 in distinct cell types of adult mouse lung. Cancer Cell *19*, 754–764. https://doi.org/10.1016/j.ccr.2011.04.019.

56. Ferone, G., Song, J.-Y., Sutherland, K.D., Bhaskaran, R., Monkhorst, K., Lambooij, J.-P., Proost, N., Gargiulo, G., and Berns, A. (2016). SOX2 Is the Determining Oncogenic Switch in Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin. Cancer Cell *30*, 519–532. https://doi.org/10.1016/j.ccell.2016.09.001.

57. Hobor, S., Al Bakir, M., Hiley, C.T., Skrzypski, M., Frankell, A.M., Bakker, B., Watkins, T.B.K., Markovets, A., Dry, J.R., Brown, A.P., et al. (2024). Mixed responses to targeted therapy driven by chromosomal instability through p53 dysfunction and genome doubling. Nat. Commun. *15*, 4871. https://doi.org/10.1038/s41467-024-47606-9.

58. Moon, Y., Lee, K.Y., Sung, S.W., and Park, J.K. (2016). Differing histopathology and prognosis in pulmonary adenocarcinoma at central and peripheral locations. J. Thorac. Dis. *8*, 169–177. https://doi.org/10.3978/j.issn.2072-1439.2016.01.15.

59. Xi, Y., Kim, T., Brumwell, A.N., Driver, I.H., Wei, Y., Tan, V., Jackson, J.R., Xu, J., Lee, D.-K., Gotts, J.E., et al. (2017). Local lung hypoxia determines epithelial fate decisions during alveolar regeneration. Nat. Cell Biol. *19*, 904–914. https://doi.org/10.1038/ncb3580.

60. Nabhan, A.N., Brownfield, D.G., Harbury, P.B., Krasnow, M.A., and Desai, T.J. (2018). Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. Science *359*, 1118–1123. https://doi.org/10.1126/science.aam6603.

61. Politi, K., Zakowski, M.F., Fan, P.-D., Schonfeld, E.A., Pao, W., and Varmus, H.E. (2006). Lung adenocarcinomas induced in mice by mutant EGF receptors found in human lung cancers respond to a tyrosine kinase inhibitor or to down-regulation of the receptors. Genes Dev. *20*, 1496–1510. https://doi.org/10.1101/gad.1417406.

62. Mookherjee, N., Ryu, M.H., Hemshekhar, M., Orach, J., Spicer, V., and Carlsten, C. (2022). Defining the effects of traffic-related air pollution on the human plasma proteome using an aptamer proteomic array: A dose-dependent increase in atherosclerosis-related proteins. Environ. Res. *209*, 112803. https://doi.org/10.1016/j.envres.2022.112803.

63. Ridker, P.M., Everett, B.M., Thuren, T., MacFadyen, J.G., Chang, W.H., Ballantyne, C., Fonseca, F., Nicolau, J., Koenig, W., Anker, S.D., et al. (2017). Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. N. Engl. J. Med. *377*, 1119–1131. https://doi.org/10.1056/NEJMoa1707914.

64. Woo, J., Lu, D., Lewandowski, A., Xu, H., Serrano, P., Healey, M., Yates, D.P., Beste, M.T., Libby, P., Ridker, P.M., et al. (2023). Effects of IL-1β inhibition on anemia and clonal hematopoiesis in the randomized CANTOS trial. Blood Adv. *7*, 7471–7484. https://doi.org/10.1182/bloodadvances.2023011578.

65. Chen, Y.-C., Hsu, C.-L., Wang, H.-M., Wu, S.-G., Chang, Y.-L., Chen, J.-S., Wu, Y.-C., Lin, Y.-T., Yang, C.-Y., Lin, M.-W., et al. (2025). Multiomics Analysis Reveals Molecular Changes during Early Progression of Precancerous Lesions to Lung Adenocarcinoma in Never-Smokers. Cancer Res. *85*, 602–617. https://doi.org/10.1158/0008-5472.CAN-24-0821.

66. Castanza, A.S., Recla, J.M., Eby, D., Thorvaldsdóttir, H., Bult, C.J., and Mesirov, J.P. (2023). Extending support for mouse data in the Molecular Signatures Database (MSigDB). Nat. Methods *20*, 1619–1620. https://doi.org/10.1038/s41592-023-02014-7.

67. Wan, J.C.M., Sasieni, P., and Rosenfeld, N. (2025). Promises and pitfalls of multi-cancer early detection using liquid biopsy tests. Nat. Rev. Clin. Oncol. *22*, 566–580. https://doi.org/10.1038/s41571-025-01033-x.

68. Takahashi, H., Ogata, H., Nishigaki, R., Broide, D.H., and Karin, M. (2010). Tobacco smoke promotes lung tumorigenesis by triggering IKKbeta- and JNK1-dependent inflammation. Cancer Cell *17*, 89–97. https://doi.org/10.1016/j.ccr.2009.12.008.

69. Weeden, C.E., Gayevskiy, V., Marceaux, C., Batey, D., Tan, T., Yokote, K., Ribera, N.T., Clatch, A., Christo, S., Teh, C.E., et al. (2023). Early immune pressure initiated by tissue-resident memory T cells sculpts tumor evolution in non-small cell lung cancer. Cancer Cell *41*, 837–852.e6. https://doi.org/10.1016/j.ccell.2023.03.019.

70. Taguchi, A., Politi, K., Pitteri, S.J., Lockwood, W.W., Faça, V.M., Kelly-Spratt, K., Wong, C.-H., Zhang, Q., Chin, A., Park, K.-S., et al. (2011). Lung cancer signatures in plasma based on proteome profiling of mouse tumor models. Cancer Cell *20*, 289–299. https://doi.org/10.1016/j.ccr.2011.08.007.

71. Taniguchi, S., Elhance, A., Van Duzer, A., Kumar, S., Leitenberger, J.J., and Oshimori, N. (2020). Tumor-initiating cells establish an IL-33-TGF-β niche signaling loop to promote cancer progression. Science *369*, eaay1813. https://doi.org/10.1126/science.aay1813.

72. Rodrigues, F.S., Karoutas, A., Ruhland, S., Rabas, N., Rizou, T., Di Blasio, S., Ferreira, R.M.M., Bridgeman, V.L., Goldstone, R., Sopena, M.L., et al. (2024). Bidirectional activation of stem-like programs between metastatic cancer and alveolar type 2 cells within the niche. Dev. Cell *59*, 2398–2413.e8. https://doi.org/10.1016/j.devcel.2024.05.020.

73. Ireland, A.S., Xie, D.A., Hawgood, S.B., Barbier, M.W., Zuo, L.Y., Hanna, B.E., Lucas-Randolph, S., Tyson, D.R., Witt, B.L., Govindan, R., et al. (2025). Basal cell of origin resolves neuroendocrine-tuft lineage plasticity in cancer. Nature *647*, 257–267. https://doi.org/10.1038/s41586-025-09503-z.

74. Panja, S., Mantri, P., Johnson, K.E., Andrade-Martinez, J.S., Yang, S.-R., Deshpande, A., Tian, H., Beg, S., Ohara, K., Leal, A., et al. (2025). Passenger mutations link cellular origin and transcriptional identity in human lung adenocarcinomas. Nat. Genet. *57*, 3066–3074. https://doi.org/10.1038/s41588-025-02418-5.

75. Gómez-López, S., Alhendi, A.S.N., Przybilla, M.J., Bordeu, I., Whiteman, Z.E., Butler, T., Rouhani, M.J., Kalinke, L., Uddin, I., Otter, K.E.J., et al. (2025). Aberrant basal cell clonal dynamics shape early lung carcinogenesis. Science *388*, eads9145. https://doi.org/10.1126/science.ads9145.

76. Mega, J.L., Stitziel, N.O., Smith, J.G., Chasman, D.I., Caulfield, M.J., Devlin, J.J., Nordio, F., Hyde, C.L., Cannon, C.P., Sacks, F.M., et al. (2015). Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. Lancet *385*, 2264–2271. https://doi.org/10.1016/S0140-6736(14)61730-X.

77. Maresso, K.C., Tsai, K.Y., Brown, P.H., Szabo, E., Lippman, S., and Hawk, E.T. (2015). Molecular cancer prevention: Current status and future directions. CA A Cancer J. Clinicians *65*, 345–383. https://doi.org/10.3322/caac.21287.

78. Choi, B., Liu, G.Y., Sheng, Q., Amancherla, K., Perry, A., Huang, X., San José Estépar, R., Ash, S.Y., Guan, W., Jacobs, D.R., et al. (2024).

Proteomic Biomarkers of Quantitative Interstitial Abnormalities in COPDGene and CARDIA Lung Study. Am. J. Respir. Crit. Care Med. *209*, 1091–1100. https://doi.org/10.1164/rccm.202307-1129OC.

79. Kyriazopoulou, E., Poulakou, G., Milionis, H., Metallidis, S., Adamis, G., Tsiakos, K., Fragkou, A., Rapti, A., Damoulari, C., Fantoni, M., et al. (2021). Early treatment of COVID-19 with anakinra guided by soluble urokinase plasminogen receptor plasma levels: a double-blind, randomized controlled phase 3 trial. Nat. Med. *27*, 1752–1760. https://doi.org/10.1038/s41591-021-01499-z.
80. Álvez, M.B., Bergström, S., Kenrick, J., Johansson, E., Åberg, M., Akyildiz, M., Altay, O., Sköld, H., Antonopoulos, K., Apostolakis, E., et al. (2025). A human pan-disease blood atlas of the circulating proteome. Science *390*, eadx2678. https://doi.org/10.1126/science.adx2678.
81. ten Haaf, K., van Rosmalen, J., and de Koning, H.J. (2015). Lung cancer detectability by test, histology, stage and gender: estimates from the NLST and the PLCO trials. Cancer Epidemiol. Biomark. Prev. *24*, 154–161. https://doi.org/10.1158/1055-9965.EPI-14-0745.
82. Spandidos, A., Wang, X., Wang, H., and Seed, B. (2010). PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. Nucleic Acids Res. *38*, D792–D799. https://doi.org/10.1093/nar/gkp1005.
83. Yang, H., Zhou, H., Feng, P., Zhou, X., Wen, H., Xie, X., Shen, H., and Zhu, X. (2010). Reduced expression of Toll-like receptor 4 inhibits human breast cancer cells proliferation and inflammatory cytokines secretion. J. Exp. Clin. Cancer Res. *29*, 92. https://doi.org/10.1186/1756-9966-29-92.
84. Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., et al. (2017). QuPath: Open source software for digital pathology image analysis. Sci. Rep. *7*, 16878. https://doi.org/10.1038/s41598-017-17204-5.
85. Wang, L., Sharma, K., Deng, H.-X., Siddique, T., Grisotti, G., Liu, E., and Roos, R.P. (2008). Restricted expression of mutant SOD1 in spinal motor neurons and interneurons induces motor neuron pathology. Neurobiol. Dis. *29*, 400–408. https://doi.org/10.1016/j.nbd.2007.10.004.
86. Marino, S., Vooijs, M., van Der Gulden, H., Jonkers, J., and Berns, A. (2000). Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. Genes Dev. *14*, 994–1004. https://doi.org/10.1101/gad.14.8.994.
87. Lim, K., Rutherford, E.N., Delpiano, L., He, P., Lin, W., Sun, D., Van den Boomen, D.J.H., Edgar, J.R., Bang, J.H., Predeus, A., et al. (2025). A novel human fetal lung-derived alveolar organoid model reveals mechanisms of surfactant protein C maturation relevant to interstitial lung disease. EMBO J. *44*, 639–664. https://doi.org/10.1038/s44318-024-00328-6.
88. Rock, J.R., Onaitis, M.W., Rawlins, E.L., Lu, Y., Clark, C.P., Xue, Y., Randell, S.H., and Hogan, B.L.M. (2009). Basal cells as stem cells of the mouse trachea and human airway epithelium. Proc. Natl. Acad. Sci. USA *106*, 12771–12775. https://doi.org/10.1073/pnas.0906850106.
89. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420. https://doi.org/10.1038/nbt.4096.
90. Li, G., Tian, L., Goodyer, W., Kort, E.J., Buikema, J.W., Xu, A., Wu, J.C., Jovinge, S., and Wu, S.M. (2019). Single cell expression analysis reveals anatomical and cell cycle-dependent transcriptional shifts during heart development. Development *146*, dev173476. https://doi.org/10.1242/dev.173476.
91. Alonso-Curbelo, D., Ho, Y.-J., Burdziak, C., Maag, J.L.V., Morris, J.P., Chandwani, R., Chen, H.-A., Tsanov, K.M., Barriga, F.M., Luan, W., et al. (2021). A gene-environment-induced epigenetic program initiates tumorigenesis. Nature *590*, 642–648. https://doi.org/10.1038/s41586-020-03147-x.
92. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
93. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.
94. Martínez-Ruiz, C., Black, J.R.M., Puttick, C., Hill, M.S., Demeulemeester, J., Larose Cadieux, E., Thol, K., Jones, T.P., Veeriah, S., Naceur-Lombardelli, C., et al. (2023). Genomic-transcriptomic evolution in lung cancer and metastasis. Nature *616*, 543–552. https://doi.org/10.1038/s41586-023-05706-4.
95. Buuren, S.V., and Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate Imputation by Chained Equations in *R*. J. Stat. Soft. *45*. https://doi.org/10.18637/jss.v045.i03.
96. Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701.
97. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proc. ACM SIGKDD Int. Conf. on Knowl. Discov. Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785.
98. Lemaitre, G., Nogueira, F., and Aridas, C.K. (2017). Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J. Mach. Learn. Res. *18*, 559–563.
99. Robbins, H.A., Alcala, K., Swerdlow, A.J., Schoemaker, M.J., Wareham, N., Travis, R.C., Crosbie, P.A.J., Callister, M., Baldwin, D.R., Landy, R., et al. (2021). Comparative performance of lung cancer risk models to define lung screening eligibility in the United Kingdom. Br. J. Cancer *124*, 2026–2034. https://doi.org/10.1038/s41416-021-01278-0.
100. Wright, J.D., Folsom, A.R., Coresh, J., Sharrett, A.R., Couper, D., Wagenknecht, L.E., Mosley, T.H., Ballantyne, C.M., Boerwinkle, E.A., Rosamond, W.D., et al. (2021). The ARIC (Atherosclerosis Risk In Communities) Study: JACC Focus Seminar 3/8. J. Am. Coll. Cardiol. *77*, 2939–2959. https://doi.org/10.1016/j.jacc.2021.04.035.
101. Joshu, C.E., Barber, J.R., Coresh, J., Couper, D.J., Mosley, T.H., Vitolins, M.Z., Butler, K.R., Nelson, H.H., Prizment, A.E., Selvin, E., et al. (2018). Enhancing the Infrastructure of the Atherosclerosis Risk in Communities (ARIC) Study for Cancer Epidemiology Research: ARIC Cancer. Cancer Epidemiol. Biomark. Prev. *27*, 295–305. https://doi.org/10.1158/1055-9965.EPI-17-0696.
102. Ru, M., Douville, C., Guenoun, A., Zahed, H., Ballantyne, C.M., Butler, K.R., Coresh, J., Couper, D.J., Galiatsatos, P., Gunter, M.J., et al. (2026). A smoking-related plasma protein score and smoking-related cancer risk and mortality in ARIC. JNCI J. Natl. Cancer Inst. *118*, 917–925. https://doi.org/10.1093/jnci/djag004.
103. Zhang, J., Dutta, D., Köttgen, A., Tin, A., Schlosser, P., Grams, M.E., Harvey, B., Yu, B., Boerwinkle, E., Coresh, J., et al. (2022). Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. Nat. Genet. *54*, 593–602. https://doi.org/10.1038/s41588-022-01051-w.
104. van Donkelaar, A., Hammer, M.S., Bindle, L., Brauer, M., Brook, J.R., Garay, M.J., Hsu, N.C., Kalashnikova, O.V., Kahn, R.A., Lee, C., et al. (2021). Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. Environ. Sci. Technol. *55*, 15287–15300. https://doi.org/10.1021/acs.est.1c05309.
105. Hammer, M.S., Van Donkelaar, A., Bindle, L., Sayer, A.M., Lee, J., Hsu, N.C., Levy, R.C., Sawyer, V., Garay, M.J., Kalashnikova, O.V., et al. (2023). Assessment of the impact of discontinuity in satellite instruments and retrievals on global PM2.5 estimates. Remote Sens. Environ. *294*, 113624. https://doi.org/10.1016/j.rse.2023.113624.
106. Durney, C.H., Tawara, A., Brauer, M., Atkar-Khattra, S., Myers, R., Meza, R., and Lam, S. (2026). APEX: A Web-Based Tool for Assessing Long-Term Outdoor PM2.5 Exposure–Brief Report. JTO Clin. Res. Rep. *7*, 100975. https://doi.org/10.1016/j.jtocrr.2026.100975.

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Human EGFR-L858R	Cell Signaling Technology	Cat# 3197; RRID:AB_1903955
Red fluorescent protein (RFP)	Rockland	Cat# 600-401-379; RRID:AB_2209751
Cldn4	ThermoFisher Scientific	Cat# 36-4800; RRID:AB_2533262
SPC	Abcam	Cat# ab211326; RRID:AB_2927746
Krt5	Abcam	Cat# ab64081; RRID:AB_1139385
Synaptophysin	Sigma-Aldrich	Cat# SAB4200544; RRID:AB_3718600
CC10	Santa Cruz	Cat# sc-130411; RRID:AB_2183388
CD68	Abcam	Cat# ab283654; RRID:AB_2922954
Anti-Mouse CD16/CD32	BD Biosciences	Cat# 553142; RRID:AB_394657
TotalSeq™-A0301 anti-mouse Hashtag 1	BioLegend	Cat# 155801; RRID:AB_2750032
TotalSeq™-A0302 anti-mouse Hashtag 2	BioLegend	Cat# 155803; RRID:AB_2750033
TotalSeq™-A0303 anti-mouse Hashtag 3	BioLegend	Cat# 155805; RRID:AB_2750034
TotalSeq™-A0304 anti-mouse Hashtag 4	BioLegend	Cat# 155807; RRID:AB_2750035
TotalSeq™-A0305 anti-mouse Hashtag 5	BioLegend	Cat# 155809; RRID:AB_2750036
TotalSeq™-A0306 anti-mouse Hashtag 6	BioLegend	Cat# 155811; RRID:AB_2750037
TotalSeq™-A0307 anti-mouse Hashtag 7	BioLegend	Cat# 155813; RRID:AB_2750039
TotalSeq™-A0308 anti-mouse Hashtag 8	BioLegend	Cat# 155815; RRID:AB_2750040
TotalSeq™-A0309 anti-mouse Hashtag 9	BioLegend	Cat# 155817; RRID:AB_2750042
TotalSeq™-A0310 anti-mouse Hashtag 10	BioLegend	Cat# 155819; RRID:AB_2750043
Anti-mouse CD45-BV421	BioLegend	Cat# 103133; RRID:AB_10899570
Anti-mouse CD31BV421	BioLegend	Cat# 102423; RRID:AB_2562186
Anti-mouse TER-119BV421	BioLegend	Cat# 116234; RRID:AB_2562917
Anti-mouse CD326 (EpCAM)-APC-Fire750	BioLegend	Cat# 118230; RRID:AB_2629758
Anti-mouse CD45.2-AF647	BioLegend	Cat# 109817; RRID:AB_492871
Anti-mouse E-cadherin-AlexaFluor647	BioLegend	Cat# 147308; RRID:AB_2563955
InVivoMAb anti-mouse/rat IL-1β	Bio X Cell	Cat# BE0246; RRID:AB_2687727
InVivo polyclonal Armenian hamster IgG	Bio X Cell	Cat# BE0091; RRID:AB_1107773
Bacterial and virus strains		
Ad5-CMV-Cre	Viral Vector Core, University of Iowa	#VVC-U of Iowa-5
Ad5-mSPC-Cre	Viral Vector Core, University of Iowa	#VVC-Berns-1168
Ad5-CC10-Cre	Viral Vector Core, University of Iowa	#VVC-Berns-1166
Ad5-CGRP-Cre	Viral Vector Core, University of Iowa	#VVC-Berns-1160
Ad5-bk5-Cre	Viral Vector Core, University of Iowa	#VVC-Berns-1547
Chemicals, peptides, and recombinant proteins		
Polidocanol	Sigma-Aldrich	Cat# P9641
Fine particulate matter (SRM2786)	Sigma-Aldrich	Cat# NIST2786
Doxycycline rodent diet	Teklad™ Custom Diets	Cat# TD.01306
DNase I	Sigma-Aldrich	Cat# D4263
Liberase™ TH	Roche	Cat# 5401135001
Liberase™ TM	Roche	Cat# 5401119001
ACK lysing buffer	Gibco	Cat# A1049201
Dispase II	Roche	Cat# 4942078001

(*Continued on next page*)

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Trypsin-EDTA	Gibco	Cat# 15400-54
LIVE/DEAD™ Fixable Aqua Dead Cell Stain Kit	ThermoFisher Scientific	Cat# L34957
LIVE/DEAD™ Fixable Green Dead Cell Stain Kit	ThermoFisher Scientific	Cat# L23101
Protector RNAse Inhibitor	Roche	Cat# 3335402001
Low melting agarose	Sigma-Aldrich	Cat# A9414
LGK974	MedChem Express	Cat# HY-17545
Recombinant human interleukin-1 beta	Merck	Cat# IL038
Tetracycline hydrochloride	Sigma-Aldrich	Cat# T7660
Penicillin-Streptomycin	Gibco	Cat# 15140122
Advanced DMEM/F12	Gibco	Cat#12634028
HEPES	Gibco	Cat# 15630080
GlutaMAX™ Supplement	Gibco	Cat# 35050061
B27 supplement (minus Vitamin A)	Gibco	Cat# 12587001
N2 supplement	Gibco	Cat# 17502001
N-Acetylcysteine	Sigma-Aldrich	Cat# A9165
Dexamethasone	Merck	Cat# D4902
cAMP	Selleckchem	Cat# S7857
IBMX	Merck	Cat# I5879
DAPT	Merck	Cat# D5942
Recombinant human FGF7	PeproTech	Cat# 100-19-50UG
CHIR99021	Tocris	Cat# 4423/10
A83-01	Focus Biomolecules	Cat# 10-1327
Geltrex™	Gibco	Cat# A14133-02
DAPI	Sigma-Aldrich	Cat# D9542
Critical commercial assays		
IL-1β RNAscope	ACD Bio-Techne	Cat# 316898
Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Bundle	10x Genomics	Cat# CG000338
Chromium Single Cell 3' Reagent Kits User Guide (v3.1 - Dual Index), with Hashing (Totalseq A)	10x Genomics	Cat# CG000315
Murine LAMP3 ELISA kit	Abbexa	Cat# abx530670
RNeasy Mini Kit	Qiagen	Cat# 74104
Luna® Universal One-Step RT-qPCR Kit	NEB	Cat# E3005
TURBO DNA-free™ Kit	Invitrogen	Cat# AM1907
Deposited data		
Murine single-nucleus RNA-seq, single-cell RNA-seq, bulk RNA-seq and proteomics data alongside processed TRACERx patient data	This paper	https://doi.org/10.5281/zenodo.19372114
UK Biobank data	UK Biobank	https://www.ukbiobank.ac.uk/
EPIC Norfolk Data	EPIC Norfolk	https://www.epic-norfolk.org.uk/
ARIC Data	ARIC	https://www5.cscc.unc.edu/aric9/
EPIC data	EPIC	https://epic.iarc.fr/
China Kadoorie Biobank Data	China Kadoorie Biobank	https://www.ckbiobank.org/
Bulk RNA-seq of pre-invasive lesions	Chen et al.[65]	EGAD50000000637
UKCTOCS Data	UKCTOCS	https://www.mrcctu.ucl.ac.uk/studies/all-studies/u/ukctocs/
LC3 Consortium Data	Albanese et al.[43]	https://lc3.iarc.who.int/
TALENT Data	TALENT	https://clinicaltrials.gov/study/NCT02611570

(*Continued on next page*)

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
CANTOS Data	Novartis	https://clinicaltrials.gov/study/NCT01327846
Human Lung Cell Atlas	Human Lung Cell Atlas	https://data.humancellatlas.org/hca-bio-networks/lung/atlases/lung-v1-0
Bulk Human RNA-Seq data	GTEx Consortium	https://gtexportal.org/home/
deCODE genetics	Eldjarn et al.[44]	https://www.decode.com/
Experimental models: Cell lines		
Foetal lung-derived human AT2 organoid line #16392	Laboratory of Emma Rawlins	https://www.rawlins.group.gurdon.cam.ac.uk/
Foetal lung-derived human AT2 organoid line #17916	Laboratory of Emma Rawlins	https://www.rawlins.group.gurdon.cam.ac.uk/
Experimental models: Organisms/strains		
Mouse: TetO-EGFR^{L858R}	National Cancer Institute	MGI:3690078
Mouse: C57BL/6J	The Francis Crick Institute	N/A
Mouse: Rosa26-tTA	Jackson Laboratories	#008600
Mouse: Rosa26-LSL-tdTomato	Jackson Laboratories	#007914
Mouse: Trp53$^{fl/fl}$	Jackson Laboratories	#008462
Mouse: CCSP-rtTA	Jackson Laboratories	#006232
Oligonucleotides		
WFDC2 forward 5'-AGAACTGCACGCAAGAGTG-3', reverse 5'-TTGAGGTTGTCGGCGCATT-3'	PrimerBank, Spandidos et al.[82]	PrimerBank ID: 56699494c1
CXCL17 forward 5'-TGCTGCCACTAATGCTGATGT-3', reverse 5'-CTCAGGAACCAATCTTTGCACT-3'	PrimerBank, Spandidos et al.[82]	PrimerBank ID: 38348269c1
LAMP3 forward 5'-GCGTCCCTGGCCGTAATTT-3', reverse 5'-TGCTTGCTTAGCTGGTTGCT-3'	PrimerBank, Spandidos et al.[82]	PrimerBank ID: 156627583c1
SFTPD, forward 5'-CCTTACAGGGACAAGTACAGCA-3', reverse 5'-CTGTGCCTCCGTAAATGGTTT-3'	PrimerBank, Spandidos et al.[82]	PrimerBank ID: 61699225c2
GAPDH, forward 5'-GGATTTGGTCGTATTGGG-3', reverse 5'-GGAAGATGGTGATGGGATT-3'	Yang et al.[83]	N/A
Software and algorithms		
R	The R Foundation	https://www.r-project.org
Python	Python Software Foundation	https://www.python.org/
Prism	GraphPad Software	https://www.graphpad.com/
QuPath	Bankhead et al.[84]	https://qupath.github.io

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Mouse models

All animal regulated procedures were approved by The Francis Crick Institute BRF Strategic Oversight Committee, incorporating the Animal Welfare and Ethical Review Body, conforming with UK Home Office guidelines and regulations under the Animals (Scientific Procedures) Act 1986, including Amendment Regulations 2012. Animals were housed in ventilated cages with unlimited access to food and water and weighed weekly. Male and female mice aged 6–15 weeks were used for functional experiments (based on availability from breeding), female mice aged 6-15 weeks were used for single-cell and single-nuclei sequencing experiments to reduce sex-based transcriptional variation. No sex-based differences in tumor formation or survival were observed. All mice were immunocompetent from C57BL/6J background from a Specific Pathogen Free (SPF) facility and naive prior to use. Daily welfare checks were carried out. Mice were genotyped (Transnetyx) and placed in groups of 1–5 mice in individually ventilated cages with a 12-hour daylight cycle.

TetO-EGFR-L858R[61] mice were obtained from the National Cancer Institute Mouse Repository. Rosa26-tTA,[85] Rosa26-LSL-tdTomato and Trp53fl/fl[86] mice were obtained from the Jackson laboratory. Mice were backcrossed onto a C57BL/6J background and further crossed as previously described[6,57] to generate Rosa26$^{LSL\text{-}tdTomato/+}$ reporter (T) mice, *Rosa26*$^{LSL\text{-}tTa/LSL\text{-}tdTomato}$*;TetO-EGFR*^{L858R} (ET) and *Rosa26*$^{LSL\text{-}tTa/LSL\text{-}tdTomato}$*;TetO-EGFR*^{L858R}*;Trp53*$^{fl/fl}$ (EPT) mice. *CCSP-rtTa;TetO-EGFR*^{L858R} mice have been described (*EGFR*-dox).[61]

Mouse lung organoids and cell lines

Organoids were grown as previously described.[6] Briefly, 2,000–10,000 sorted tdTomato+ live cells were resuspended in 3D organoid medium consisting of DMEM/F12 with 10% FBS, 100 U/mL penicillin–streptomycin, 1x Insulin-Transferrin-Selenium, and 1 mM l-glutamine (all from Gibco) and 1 mM HEPES (in-house). Cells were mixed at a 1:5 ratio with the MLg2908 mouse lung fibroblast cell line (ATCC, Cat# CCL-206), provided by the Cell Services Unit of The Francis Crick Institute where cells were routinely screened for mycoplasma and authenticated using short-tandem repeat profiling, and maintained in DMEM containing 10% FBS and 100 U/mL penicillin-streptomycin at 37°C and 5% CO_2 at low passage numbers. The cell mixture was resuspended in growth-factor-reduced Matrigel (Corning, Cat#356231) at a 1:1 ratio and 100 μL pipetted into a 24-well Transwell insert with a 0.4 μm pore (Corning, Cat#3470). Organoids were cultured in 3D organoid medium and counted after 14 days using an EVOS microscope (ThermoFisher Scientific).

Murine lung organoid dataset

Single-cell RNA-seq data from murine lung organoids were accessed from Choi et al.[21] (GSE144468); experimental model details, including mouse strain, age, and husbandry conditions, are described in the original publication.[21]

UK Biobank

The UK Biobank (UKBB) is a prospective cohort study that recruited *N* = 502,401 participants, aged 37–73, from 2006–2010, with a subset of individuals (*N* = 54,219) having Olink® plasma proteomics measured (for 2,923 proteins) from baseline blood samples alongside sex and ancestry data, as previously described.[27] UKBB data were accessed under project number 82693 and ethical approval given by the North West Multicentre Research Ethics Committee, the National Information Governance Board for Health and Social Care and the Community Health Index Advisory Group. The UKBB determined cancer incidence through linkage with national cancer registries with diagnoses recorded using the tenth revision of the International Classification of Diseases (ICD10) codes. Participants were excluded according to the following criteria: any cancer diagnosed pre-recruitment, a cancer diagnosis date entry but no corresponding cancer annotation. No formal power calculation was performed; the full available cohort with baseline Olink® plasma proteomics data were used (*N* = 48,099).

External validation datasets

Eight external validation datasets were used (sex distribution and cohort characteristics are detailed in Table S1): the European Prospective Investigation into Cancer and Nutrition study (EPIC), EPIC-Norfolk (analysed in two separate batches, as per published recommendations[45]), the Atherosclerosis Risk in Communities study (ARIC), the China Kadoorie Biobank, the LC3 consortium study, the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS), and a study led by deCODE genetics, with details on race and ethnicity available from each cohort publication.[42–47] Data access was obtained for EPIC-Norfolk, UKCTOCS, and China Kadoorie Biobank, with summary statistics utilised for the LC3 consortium and deCODE genetics from two previously reported manuscripts.[43,44] Within each cohort, plasma protein levels were measured; however, due to limitations with relative quantification of the assay, direct comparison of proteomic measurements between cohorts was not possible. All participants provided written informed consent, and ethical approval was obtained from the relevant institutional review boards: the International Agency for Research on Cancer ethics committee (IEC 22-01) for EPIC; the Norfolk Research Ethics Committee (05/Q0101/191) for EPIC-Norfolk; the institutional review board at each participating centre for ARIC; the UK North West Multicentre Research Ethics Committee (00/8/34) for UKCTOCS; and the Oxford Tropical Research Ethics Committee, the Chinese Center for Disease Control and Prevention, the Chinese Academy of Medical Sciences, and the Peking University Institutional Review Board for CKB. Ethics approvals for the LC3 consortium and deCODE genetics cohorts are described in their respective manuscripts.[43,44] Sex, age, and smoking status were adjusted for in relevant analyses where available; formal sex- or ancestry-stratified analyses of signature performance were not performed as the study was not powered for such comparisons.

TRACERx study

The TRACERx observational study (NCT01888601) has approval from the UK research and ethics committee (13/LO/1546). Plasma proteomics data were collected from individuals prior to surgical resection and at follow-up from individuals who did not relapse within at least two years post-surgery, with detailed demographic data previously described,[53] and smoking status reported via a questionnaire at study entry. No formal power calculation was performed; the full available cohort with plasma proteomics data were used (*N* = 482).

CANTOS study

The Canakinumab Anti-inflammatory Thrombosis Outcome Study (CANTOS; NCT01327846) enrolled 10,061 individuals with a history of myocardial infarction and baseline levels of high-sensitivity C-reactive protein (hsCRP) $\geq$ 2 mg/L.[63] The CANTOS study was conducted from April 2011 to June 2017 across 39 countries and sponsored by Novartis AG (Basel, Switzerland), with individuals followed up for a median of 3.7 years. Subsequent exploratory analysis found a dose-dependent reduction in lung cancer incidence between canakinumab and placebo.[7] All individuals provided written informed consent, including a separate informed consent for individuals who provided additional blood samples for genetic analyses and direct biomarker analysis. Proteomic analysis was

conducted in serum samples at these time points from the individuals who also agreed to the additional research, with sex distribution and cohort characteristics as previously described.[64] Baseline serum samples were collected from all biomarker sub-study participants and profiled using the SomaScan® v3 assay, which quantifies 4,785 unique proteins.[64] No formal power calculation was performed for the biomarker sub-study; the full available cohort with proteomic data was used.

TALENT study

The TALENT (NCT02611570) study was a multi-centre prospective clinical trial conducted across 17 tertiary medical centres in Taiwan between 2015 and 2019.[54] Written informed consent was obtained from all participants and the study protocol was approved by the individual institutional review boards. All participants screened negative for lung cancer via chest X-ray at study entry, with plasma samples collected at enrolment. Subsequent incident diagnoses of invasive lung adenocarcinoma were identified through linkage with the cancer registry. Proteomic analysis was performed on baseline plasma samples from a subset of TALENT participants (n = 251 cases and n = 501 controls; 81.3% female). Controls were determined by a 1:2 matched case-control design, with controls selected by matching on age, sex, and baseline smoking status. Assuming a two-sided α of 0.05 and 80% power, the study was powered to detect a modest association between protein levels and lung cancer risk (OR ~1.25–1.30 per SD increase). Four Olink® Target 96 panels (Oncology III, Oncology II, Cardiovascular III, and Immune Oncology) were assayed per sample; not all 14 proteins of interest were covered across these panels due to resource constraints.

Air pollution exposure study

The controlled diesel exhaust exposure study was approved by the institutional ethics review board of the University of British Columbia and the Vancouver Coastal Health Research Institute (H14-00821 and H16-03053). The participants were enrolled with written informed consent, and the investigation conformed to the principles outlined in the Declaration of Helsinki. The participants were 4 males, and 2 females, aged 40-66 years, and sample size was determined based on previous studies by the researchers.[62]

Bulk human pre-invasive RNA-seq

Bulk RNA-seq data from human pre-invasive lung lesions were accessed from Chen et al.[65] through the European Genome-Phenome archive (EGAD50000000637); participant characteristics, ethics approval, and informed consent procedures are described in the original publication.[65]

Human foetal lung-derived alveolar organoids

Foetal lung-derived alveolar organoids (from the Rawlins lab, University of Cambridge) were cultured as Geltrex-domes (Gibco) in AT2 medium as described in Lim et al. (2025).[87] Tissue was obtained from the Human Developmental Biology Resource (HDBR) with written informed donor consent; donors were informed of the purpose of the research and the specific procedures for which the tissue would be used. Organoids were derived from foetal lung tissue at 17 and 20 weeks, from male donors. Development of the Human Developmental Biology Resource was reviewed by the North East Ethics Committee (23/NE/0135: IRAS project ID: 330783) and specifically for this project at University College London (REC reference: 23/LO/0312: IRAS project ID: 326492). This research was conducted in accordance with UK HTA regulations.

METHOD DETAILS

Animal procedures

Mice were randomly assigned to experimental groups with group size based on prior work with the model. Cre-mediated recombination was initiated by intratracheal delivery of adenoviral CMV-Cre (2.5×10^7 virus particles per 50 μL), by Ad5-bk5-Cre, Ad5-CGRP-Cre, Ad5-CC10-Cre or Ad5-SPC-Cre (2.5×10^8 virus particles per 50 μL donated by A. Berns,[13,55,56] Viral Vector Core, University of Iowa) or by using chow containing doxycycline (Harlan Teklad). For single-nuclei sequencing studies, 1.0×10^{10} virus particles per 50 μL were used to isolate sufficient nuclei for sequencing. Mice were pre-treated with intratracheal delivery of 2% polidocanol (13.5 μL, Sigma-Aldrich) before Ad5-bk5-Cre, Ad5-CMV-Cre or PBS control treatment. During survival studies, mice were weighed weekly and were euthanised when the humane endpoint of 15% weight loss from baseline was reached or any sign of distress was observed (i.e. hunched, piloerection, difficulty breathing). *In vivo* PM exposure was modelled by intratracheal administration of 50 μg SRM2786 (Sigma-Aldrich) in 50 μL, or PBS control, three times a week for three weeks as previously described.[6]

MicroCT imaging

EPT mice were anaesthetised with isoflurane and subject to monthly thorax scanning using a Quantum GX2 microCT scanner (Perkin Elmer). Nodules were called observable when measured at a maximum diameter of 0.3mm in axial images and observed to increase in size in the next consecutive microCT scan. 3D reconstruction of lungs was performed using AnalyzeDirect (USA).

Histopathology and immuno-staining

Mouse lung and trachea were perfused with 10% formalin and fixed overnight in 10% formalin before embedding in paraffin wax. 4μm tissue sections were cut, deparaffinized and rehydrated using standard methods. Sections were either routinely stained with H&E or,

for immunohistochemistry, underwent antigen retrieval using pH 6.0 citrate buffer and incubated with antibodies: human EGFR-L858R (Cell Signaling Technology, #3197), RFP (Rockland, 600-401-379), Cldn4 (ThermoFisher Scientific, 36-4800), SPC (Abcam, ab211326), Krt5 (Abcam, ab64081), synaptophysin (Merck, SAB4200544), CC10 (Santa Cruz, sc-130411). Primary antibodies were detected either using biotinylated secondary antibodies, followed by HRP or DAB, or with subsequent Opal fluorescence secondary antibodies (Akoya). A commercial kit was used to detect IL-1β RNA transcripts by RNAscope (ACD Bio-Techne) following the manufacturer's instructions. Labelling for CD68 protein was subsequently performed and detected using Opal fluorescence following the manufacturer's protocols (Akoya). Probes visualised through fluorescence were used to detect IL-1β RNA and CD68 protein simultaneously. Slides were imaged using a Zeiss AxioScan.Z1 slide scanner or a Polaris slide scanner with unmixing for staining with 4 channels. Tumor grading and lesion analysis were carried out by board-certified veterinary pathologists. Image analysis was carried out using QuPath.[84] Histopathology analyses were performed blinded to genotype and treatment condition.

Flow cytometry

To generate lung single cell suspensions, lung tissue was minced, then digested with Liberase TM and TH (75 μg/mL each, Roche Diagnostics) and DNase I (25 μg/mL, Sigma-Aldrich) in HBSS for 30 minutes at 37°C in a shaker at 180 r.p.m. Samples were passed through a 100 μm filter, centrifuged (300g, 5 minutes, 4°C) and red blood cells were lysed with ACK buffer (Life Technologies) following manufacturer's instructions. Tracheal single cell suspensions were generated as previously described.[88] Briefly, tracheas were incubated in dispase (16 U/mL, Roche) for 40 minutes at room temperature, the reaction stopped with 5% FBS in DMEM (Gibco), followed by peeling epithelium with forceps. Epithelial sheets were washed with PBS and incubated with trypsin-EDTA (1X, Gibco) for 20 minutes at 37°C before quenching with 5% FBS-DMEM and filtering through a 100 μm filter. Extracellular antibody staining was performed with viability determined by Live/Dead Aqua (ThermoFisher Scientific) or DAPI (Sigma-Aldrich) staining as described below. Cell sorting was performed on Influx, Aria Fusion or Aria III instruments (BD Biosciences).

Single-nuclei RNA sequencing

Lung single cell suspensions were stained with Live/Dead Aqua (ThermoFisher), CD45-BV421, CD31-BV421, Ter119-BV421 (lineage) and EpCAM-APC-Fire750; lineage-negative, EpCAM$^+$, tdTomato$^+$ live cells were sorted into collection buffer (0.04% BSA-PBS) containing Protector RNase Inhibitor (1 U/μL, Roche). Tracheal single cell suspensions were stained only with Live/Dead Aqua, and tdTomato$^+$ live cells were sorted into collection buffer as above. Nuclei from the sorted cells were isolated according to the CG00365 protocol (10x Genomics) for low cell input nuclei isolation, before proceeding with multiome-seq using the Chromium Single Cell Epi Multiome ATAC + Gene Expression protocol (10x Genomics). In total, nuclei from the following conditions were subject to multiome-seq: T mice infected with Ad5-CC10-Cre at 3 weeks (lungs from 10 mice pooled), Ad5-bk5-Cre at 10 weeks (lungs and tracheas from 10 mice pooled), Ad5-SPC-Cre at 3 weeks treated with PBS control or PM (lungs from 20 mice pooled for PBS and 10 mice for PM); ET mice infected with Ad5-bk5-Cre at 10 and 20 weeks (lungs and tracheas from 10 mice pooled per time point), Ad5-CC10-Cre at 3 and 10 weeks (lungs from 10 mice pooled per time point), Ad5-SPC-Cre at 3 and 10 weeks treated with PBS control or PM (lungs from 10 mice pooled per condition), EPT mice infected with Ad5-SPC-Cre at 25 weeks (tumors from 2 mice analysed separately) were used as a late stage LUAD reference. Data from the Ad5-SPC-Cre-induced PBS control cells were used both in cell-of-origin (Figure 2) and PM exposure analyses (Figure 4) in this manuscript.

snRNA-seq preprocessing

Raw sequencing data were processed using CellRanger-ARC (v2.0.1) with a custom reference containing the mm10 (GENCODE vM23/ Ensembl 98) genome released by 10x Genomics (2020–a), the entire *TdTomato* viral insert sequence and the *EGFR*^{L858R} sequence.

snRNA-seq QC, clustering and annotation

QC was performed, per sample, in R (v4.3.2) with a randomly generated seed using Seurat (v4.4.0).[89] Doublets were inferred and removed using DoubletFinder (v2.0.4). Cells passing the following filters were retained: 200 < nFeature < 7500, mitochondrial gene content < 20%, ribosomal gene content < 20%, haemoglobin gene content < 10%, and platelet gene content < 10%. Normalisation, scaling, and variable feature selection was performed using SCTransform(vst.flavor = 'v2'), regressing out cell cycle genes[90] and mitochondrial, ribosomal, haemoglobin, and platelet gene content. Contaminating non-epithelial cells were removed. Downstream analysis was performed with a set seed. Seurat's FindAllMarkers() was used to identify cell type-specific marker genes and FindMarkers() for differential expression analysis; adjusted p-values were calculated using the Bonferroni method. Differentially upregulated genes of KAC-like states from prior studies[16,17,20,21] were used as signatures (using Seurat's AddModuleScore() function) to support identification of KAC clusters in our dataset. Gene sets overlapping with LUAD were determined based on methods in Alonso-Curbelo et al.[91] with wild-type AT2 cells from PBS-treated wild-type control mice used as controls and adenocarcinomas from EPT mice induced with AT2-restricted virus used as a reference for late-stage LUAD. Collective expression of transcripts encoding the 14 plasma proteins of interest was quantified as an aggregated signature using AddModuleScore().

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

Single-cell RNA sequencing (scRNA-seq)

Lungs were profiled from T mice infected with Ad5-SPC-Cre (resulting in rare tdTomato$^+$ wild-type cells) followed by three weeks treatment with PBS control or PM three times a week by intra-tracheal intubation and collected acutely after the final treatment; ET mice infected with Ad5-SPC-Cre followed by treatment with PBS control or PM (resulting in rare *EGFR* mutant cells, lungs from 10 mice pooled per condition). T mice treated with PBS were used as healthy control mice analysed in Figure 1. Lung single cell suspensions from individual mice (n = 10 per group) were first stained with Totalseq A hashtag antibodies (Biolegend) and then combined for bulk processing within each condition. Cells were stained with Live/Dead Green (ThermoFisher), CD31-BV421, Ter119-BV421, CD45.2-AF647 and EpCAM-APC-Fire750 antibodies. Immune cells (live CD45$^+$), tdTomato$^-$ epithelial cells (live CD45- lineage- EpCAM$^+$ tdTomato-), tdTomato$^+$ epithelial cells (live CD45- lineage- EpCAM$^+$ tdTomato+) and other lung cells (live CD45- EpCAM-) were sorted separately into collection buffer (10% FBS-PBS). Fixed proportions of sorted cell populations (60% immune cells, 19% tdTomato- epithelial cells, 1% tdTomato$^+$ epithelial cells and 20% other lung cells) were combined for scRNA-seq within each condition using the 10x Genomics 3' Gene Expression kit.

scRNA-seq preprocessing

Raw sequencing data were processed using CellRanger (v7.1.0) with a custom reference containing the mm10 (GENCODE vM23/Ensembl 98) genome released by 10x Genomics (2020-A), the entire *TdTomato* viral insert sequence, and the human *EGFR*^{L858R} sequence. Cells were demultiplexed using 'cellranger multi' with default parameters.

scRNA-seq QC, clustering and annotation

QC was performed, per sample, in R (v4.3.2) with a randomly generated seed using Seurat (v4.4.0). Doublets were removed using mouse hashing demultiplexing. Cells that passed filters as above for snRNA-seq were kept, and all normalisation, scaling and variable feature selection performed as above. Cells expressing the human *EGFR*^{L858R} transgene were rare in this dataset and were filtered out to streamline analyses of the lung microenvironment cells. One mouse from the WT PBS group was excluded before analysis as its cellular composition was substantially divergent from that of other mice in the same group. Lung cell populations were annotated based on expression of cell type-specific markers, in line with prior studies.[17,20] Expression of transcripts encoding the 14 plasma proteins of interest was quantified as an aggregated signature score using AddModuleScore(). Pseudobulk analysis was performed using AggregateExpression() based on mouse hashtag IDs, and differential expression analysis was carried out using DESeq2.[92]

Analysis of murine lung organoid scRNA-seq

Published scRNA-seq data were accessed from Choi et al.[21] (GSE144468), and processed as described above. Doublets were removed using DoubletFinder (v2.0.4) and normalisation was performed using SCTransform(vst.flavor = 'v2') with default parameters. Samples were integrated for visualisation and clustering following the SCTransform integration workflow. Marker genes from the original manuscript were used to distinguish the different cell populations. Differential expression between treatment (IL-1β) and (PBS) control, within each cell population, was performed using Seurat's FindMarkers function.

PCLS generation for live cell imaging

Precision cut lung slices were generated according to an adapted protocol.[25] Mice were culled via overdose pentobarbital injection at the indicated timepoints after adenoviral Cre induction (between 6 and 16 weeks post-induction) and a catheter was inserted into the trachea. 1-3 mL of 2% low melting agarose (Sigma-Aldrich) solutions were injected through the catheter before animals were moved to ice and lungs dissected once agarose was set. Individual lung lobes were further embedded in 2% agarose before 300μm lung sections were cut using a vibratome (Leica VT1200S). Slices were stained with E-cadherin-AlexaFluor647 (Biolegend) for 1 hour at 37°C in DMEM. Slices were mounted using 2% agarose into 24 well glass bottomed plates (Ibidi) and cultured in DMEM supplemented with 1% penicillin/streptomycin (Gibco) at 37°C. For *ex vivo* Wnt inhibition, slices were treated with DMSO or 100nM LGK974. Live cell imaging was performed using 3D confocal microscopy (Olympus FV3000) for 72 hours. After 72 hours of culture, tissue slices were fixed in 4% PFA and stained for SPC (Abcam, ab211326) and imaged by confocal microscopy (Olympus FV3000). Data were analysed using Fiji.

Ex vivo PM challenge of PCLS

Lungs from *EGFR*-dox mice fed with a doxycycline-containing diet for 5 days were used to generate PCLS as described above with wild-type mice used as controls, with minor protocol modifications. Specifically, a biopsy puncher was used to create 6mm tissue cores from agarose-inflated lungs, and these were cut into 400 μm slices. PCLS were cultured for 72 hours in DMEM medium supplemented with 10% heat-inactivated fetal bovine serum, 1% penicillin/streptomycin (Gibco) and 2.5 μg/mL tetracycline, with or without 100 μg/mL of PM (SRM2786; Sigma-Aldrich). Where indicated, 100 μg/mL anti-IL-1β (InVivoMAb, BE0246) or IgG control (InVivoMAb, BE0091) were added to the media at the start of the culture. For protein secretion analysis, following 72 hours of culture supernatants from PCLS culture media were collected, centrifuged at 1,000 g for 10 minutes at 4°C, supernatants aliquoted and stored at -80°C. LAMP3 protein was measured using mouse LAMP3 ELISA Kit (Abbexa) following the manufacturer's instructions

and read on a Safire II plate reader (Tecan). For Cldn4^{+} cell abundance analysis, after 72 hours of culture, tissue slices were fixed in 4% PFA, stained using anti-Cldn4 antibodies (ThermoFisher Scientific, 36-4800) and imaged by confocal microscopy (Olympus FV3000). Data were analysed using Fiji and QuPath.

Bulk RNA-seq of murine tumor tissue

RNA was extracted from flash-frozen, dissected large single tumors collected from EPT mice at ethical endpoint (approx. 2mm^3) using the AllPrep DNA/RNA Mini Kit (Qiagen). Tissue was homogenised in Buffer RLT Plus containing β-Mercaptoethanol (Sigma-Aldrich M3148) using a fresh TissueRuptor Disposable Probe (Qiagen: 990890). The lysate was processed through a QIAshredder column (Qiagen: 79656). RNA was treated with RNase-free DNase on-column (Qiagen: 79254), eluted and stored in RNase/DNase-free water at -80°C. RNA quantity and integrity were assessed using the Qubit RNA BR assay kit (Invitrogen Q10211) and a BioAnalyser. Bulk RNA sequencing (RNA-seq) was performed on samples to yield 25 million paired-end reads per sample (PE100). Libraries were prepared using NEB mRNA polyA selection. The resulting FASTQ files were processed using Kallisto[93] (v0.45.0) with the *Mus musculus* GRCm38 reference genome, using default parameters. Differential expression analyses were conducted in R (v4.2.3) using DESeq2 (v1.38.3).

Bulk RNA-seq analysis of pre-invasive dataset

Data from Chen et al.[65] were processed using in-house pipelines.[94] Briefly, Illumina adapters were trimmed from raw sequencing reads using Cutadapt, and FASTQ files with less than 80% of total reads being duplicates were kept for alignment. FASTQs were aligned to hg38 human reference genome build using STAR (v.2.5.2a) in two-pass mode with ENCODE 3 parameters. The same reads were also mapped to the human transcriptome (GENCODE v42) using the same STAR parameters to generate gene expression data. RSEM (v.1.3.3) was used to quantify gene expression from the aligned files to the transcriptome. Samples in which less than 75% of protein-coding genes were expressed were excluded from analyses.

Machine learning model development

To predict incident lung cancer diagnoses from baseline data, we trained a machine learning classification model using 2,923 plasma proteins measured by the Olink® platform alongside patient characteristics as candidate predictors. The UKBB dataset was split into train (75%) and held-out test (25%) sets, stratifying by smoking status, sex, household income, educational attainment, lung cancer diagnosis (inferred using the ICD-10 code C34), age at baseline, body-mass index (BMI) and pack years of smoking. The latter three continuous variables were first categorised into quartiles. Missing values were treated as a separate category for the purposes of splitting, to allow the distribution of missingness across these variables to be factored into the selection of train and held-out samples. Splitting was performed using the MultilabelStratifiedShuffleSplit() function from the iterative-stratification (v0.1.7) Python package. Missing covariate data were imputed separately in the train and held-out sets to minimise data leakage, using multiple imputation with chained equations (MICE[95]; using the mice package v3.17.0). Imputed covariables were smoking status (categorised into never, previous, and current; < 1% missing), passive smoking (weekly hours of home tobacco exposure; 10.0% missing), pack-years of smoking (15.4% missing), BMI (< 1% missing), household income (dichotomised into < and >= £31,000 annually; 14.6% missing) and educational attainment (split by degree or professional qualification status; 1.3% missing). To predict values for missing data points, MICE imputation models incorporated these variables, lung cancer diagnosis, and follow-up duration in addition to: PM exposure (derived from a land-use regression model), age at baseline and sex. Imputation models were trained using the training dataset only and subsequently applied to the held-out test set. Continuous variables were imputed with predictive mean matching, while random forest and logistic regression were used to impute higher order categorical and binary variables, respectively. This produced 15 variant imputed datasets, where missing covariate data were replaced with imputed values (thus yielding complete datasets). Due to variation in the modelling process, these 15 complete datasets naturally contain small variations across imputed variables. Each imputed dataset was independently used in the same analysis protocol.

Recursive feature elimination (Probatus, v3.1.2) was used to select proteins (from the full list of 2,923 markers, excluding proteins with >25% missing values). Within the training data, the dataset was stratified by the number of incident lung cancer diagnoses and split into five folds. Bayesian hyperparameter optimization included all features with five repeats of five-fold cross-validation and was performed using the Tree-structured Parzen (TPE) sampler from the Optuna python package (v3.5.0[96]), with 100 trials. Each of these 100 trials gave a set of sampled hyperparameters to produce a trained model, with each feature given a SHAP (SHapley Additive exPlanations) score, which measures the contribution of that feature to the model prediction. After each trial, the features with the lowest 20% of SHAP scores were eliminated. Features that culminated in the peak cross-validation test set performance were selected.

Following feature selection, 100 models were created using different hyperparameter combinations, for hyperparameter tuning of the eXtreme Gradient Boost classification model (XGBoost, v2.0.3[97]), using the area under the receiver-operator characteristic (ROC–AUC) as the objective function. Due to data imbalance (N = 375 participants diagnosed with lung cancer during follow-up, compared to N = 47,724 without a lung cancer diagnosis), we randomly undersampled the majority class to generate a 1:1 ratio between cases and controls in the cross-validated training set (using imbalanced-learn v0.12.0[98]). We utilised a bagging procedure to aggregate predicted probabilities across the 100 cross-validation fold models (by the mean) to output the final predicted probability per individual.

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

Model Validation

The final model, consisting of 14 proteins alongside age, smoking status, pack-years, and past diagnosis of COPD, was benchmarked against probabilistic lung cancer risk prediction models in the UK Biobank, using the lcmodels package (v4.1.1) set out in a previous manuscript.[99] This comparison was conducted on the held-out test set (comprising 25% of the full dataset) with the original class balance (i.e. no undersampling was performed). The ROC–AUC was calculated and the statistical significance of differences in AUC values was assessed using DeLong's test (pROC, v1.18.5). For sensitivity analysis, the held-out test set was further stratified by two-year intervals prior to lung cancer diagnosis to evaluate model performance.

Protein measurement in validation cohorts

The association between each protein and lung cancer incidence was quantified using hazard ratios derived from Cox proportional hazards models, adjusting for participants' baseline age where available (see cohort-specific details below), apart from the deCODE genetics and LC3 consortium cohorts where logistic regression and conditional logistic regression were used respectively to derive odds ratios. To account for heterogeneity in cohort design and characteristics (Table S1), a random-effects meta-analysis was employed to integrate findings across cohorts. Random-effects meta-analysis was performed using the metafor R package (v 5.0-1).

UKCTOCS

Temporal patterns of serum protein levels were analysed longitudinally in a cohort of 248 women (98 lung cancer cases and 150 controls) over five years preceding clinical lung cancer diagnosis from UKCTOCS.[42] Serum samples from each individual were analysed annually using the Olink® Oncology II proteomics panel. For visualisation, we used locally estimated scatterplot smoothing (LOESS) curves to assess trends in protein levels over time relative to lung cancer diagnosis. Protein levels were standardised to have a mean of zero and a standard deviation of one prior to analysis. LOESS curves were fitted using the loess() function in R with default span. Cox regression models were used to estimate hazard ratios per standard deviation of protein values and were adjusted for age at sampling.

LC3 Consortium

The LC3 Consortium dataset comprised 731 lung cancer cases and 731 smoking-matched controls drawn from six prospective cohorts, with cases and controls matched on age, sex, smoking status, and date of inclusion.[43] We used summary statistics from Data S4 of the published manuscript,[43] in which odds ratios per standard deviation increment in relative protein concentrations were derived using conditional logistic regression.

deCODE genetics + Icelandic Cancer Project

Data from the deCODE + Icelandic Cancer Project was obtained following personal correspondence with the authors of the manuscript.[48] Associations were estimated using a logistic regression model using the SomaScan® v4 assay data collected during this project. This cohort consisted of 610 (72.4%) incident lung cancer cases, 232 (27.6%) prevalent lung cancer cases, and 37,892 non-cancer controls. Logistic regression was used to calculate odds ratios by modelling case status (defined as incident and prevalent lung cancer cases combined) versus non-lung cancer controls. Time from sample collection to lung cancer diagnosis was defined using only incident lung cancer cases.

EPIC

Results for the EPIC study were obtained through a collaboration between the Francis Crick Institute (T.P., M.A, C.S) and EPIC investigators (K.S.B, D.C.M, M.G, R.C.H.V, M.D.C, H.Z, P.M.K). Hazard ratios were adjusted for cigarettes smoked per day, number of years smoked, BMI, and educational attainment. Baseline hazards were also stratified by age (binned into 5 year age groups), sex, and recruitment centre, with proteomics assessed using the SomaScan® v4.0 platform. 25.7% (N = 188) of incident lung cancer cases in the LC3 consortium were drawn from the EPIC cohort but were assayed using Olink® panels rather than SomaScan® 7K and are therefore treated as independent observations in EPIC.

EPIC-Norfolk

Data access was obtained through an agreement between the Francis Crick Institute (T.P, C.S) and the MRC Epidemiology Unit, University of Cambridge. EPIC-Norfolk forms a subset of the EPIC study and accounts for 33.7% of incident lung cancer cases in EPIC; as EPIC-Norfolk samples were assayed using Olink® panels rather than the SomaScan® 7K platform used for the full EPIC cohort, they are reported separately. Serum samples from baseline assessment were assayed in two independent batches: batch 1 comprised a randomly selected control sub-cohort (N = 749) and case sub-cohort (N = 291), profiled using the Olink® Explore 1536 platform; batch 2 comprised a control sub-cohort (N = 1,010) and case sub-cohort (N = 698), profiled using the Olink® Explore Expansion platform. Normalised proteomic expression (NPX) values were not compared between batches owing to the absence of bridging samples. Cox regression models were used to estimate hazard ratios per standard deviation of protein values and were adjusted for age at sampling.

Please cite this article in press as: Pandya et al., Plasma signals of lung tumor promotion for molecular cancer prevention, Cell (2026), https://doi.org/10.1016/j.cell.2026.05.005

China Kadoorie Biobank

Results were obtained through collaboration between the Francis Crick Institute (T.P, C.S) and China Kadoorie Biobank investigators (N.W, K.H.C, Z.C). Proteomics data were assayed using four Olink® panels spanning 2,941 proteins from 2,029 participants (*N*= 31 cancer cases) from a sub-cohort of the China Kadoorie Biobank.[46] 13/2029 participants reported a history of cancer at baseline and were thus excluded. 3/2029 participants without Olink® data were excluded. Cox regression models were used to estimate hazard ratios per standard deviation of protein values in the sub-cohort, stratified by sex and region, and adjusted for age (plus squared term), time since last meal (plus squared term), ambient temperature (plus squared term) and educational attainment.

Atherosclerosis Risk in Communities (ARIC) Study

Results were obtained through collaboration between the Francis Crick Institute (T.P, M.A, C.S) and ARIC investigators (E.A.P, V.A.B, N.C, Z.W).[100,101] Proteomics (SomaScan® 5k) were run on the entire eligible cohort of ARIC. A Cox regression model was used to estimate hazard ratios per doubling of relative fluorescence unit (each protein was $\log_2$ transformed) adjusting for age, recruitment centre, sex, smoking status (current, former, or never), pack years smoked, BMI, waist-to-hip ratio, diabetes status, height, physical activity, as well as work, leisure, and sport indexes, alcohol drinking status, a plasma protein-derived smoking score,[102] PEER factors (Probabilistic Estimation of Expression Residuals),[103] and genetic principal components.

TALENT

Samples were processed locally in Taiwan and quality control followed an established protocol[27]: samples with QC warnings were excluded, as were samples whose median NPX exceeded ±5 standard deviations from the median NPX across all samples. Proteins with >50% of measurements below the plate-specific limit of detection were removed. $PM_{2.5}$ exposure was estimated using the Air Pollution Exposure Tool (APEX), which linked participants' residential postcodes for the year preceding enrolment with annual satellite-derived air pollution estimates at 0.01°x0.01° resolution, which were obtained from the Atmospheric Composition Analysis Group.[104–106] For each geocoded address, APEX averages all $PM_{2.5}$ grid cells within an 11-km radius to generate the final annual exposure estimate. To understand the relationship between $PM_{2.5}$ and protein levels, adjustments were made for participants' age, BMI, family history of lung cancer and sex.

GSVA analysis of TRACERx plasma proteomics

Gene Set Variation Analysis (GSVA) was applied to normalised protein expression (NPX, measured using the Olink® platform) data to enable pathway-level comparisons across cancer stages. GSVA was performed using the GSVA R package (v3.23) with default parameters and the kcdf= 'Gaussian' kernel.

Air Pollution Exposure in Humans

Plasma samples obtained from healthy participants (n=6) were collected 24 hours after exposure to 2 hours of diesel exhaust (300 $\mu g/m^3$) or filtered air (crossover design, separated by one month).[62] Plasma (70 µL each) was probed using a SOMAmer®-based proteomic array at the Manitoba Centre for Proteomics and Systems Biology. SomaScan® v1.3 was used for measuring the abundance of 1,307 distinct proteins. Protein expression profile was measured with Relative Fluorescence Unit (RFU) readouts. RFU values obtained from the proteomic arrays were $\log_2$ transformed and Welch T-test was used for differential analysis to determine the expression profile of plasma proteins following exposure to diesel exhaust compared to control. Proteins with fold change $\geq$1.5 with $p \leq 0.05$ were considered to be significantly differentially altered.

Analysis of GTEx Consortium bulk RNA-seq

We accessed bulk RNA sequencing (RNA-seq) data from up to 54 non-diseased tissue sites collected from 946 deceased individuals, as provided by the GTEx Consortium.[50] To assess whether lung tissues exhibited higher transcript expression (measured as transcripts per million, TPM) compared to other tissues, GSVA was applied and a Wilcoxon test was conducted to compare the distribution of expression values in lung tissues against that of the next highest non-lung organ.

Analysis of the Human Lung Cell Atlas data

The RDS object of the integrated cell atlas of the human lung in health and disease (core) was downloaded from the human cell atlas data portal (https://data.humancellatlas.org/hca-bio-networks/lung/atlases/lung-v1-0). Data were subsampled to 50,000 cells and ann_level_3 annotations were used for cell type information with smooth muscle FAM83D+ cells re-annotated as fibromyocytes. Normalised gene counts were used to calculate the mean expression per cell type of the 14 genes encoding the 14-protein signature proteins (*ALPP, GDF15, CXCL17, CEACAM5, WFDC2, TNFSF13B, LAMP3, SFTPA1, SFTPD, CDCP1, PLAUR, PRSS8, PIGR, MMP12*). Seurat's AddModuleScore() was used to calculate the average expression levels of the 14-protein signature on a single cell level.

Analysis of the CANTOS trial

For each individual in the study, a proteomic signature score was calculated as the mean of their baseline protein relative fluorescence units (RFU) for 10/14 aptamers available. Subjects were dichotomised by cohort-median signature score into "Higher" and

"Lower" groups. Cox proportional-hazards models estimated the association between the binarised or continuous signature score and incident lung cancer, adjusting for baseline BMI, smoking status, and age. The number needed to treat (NNT) was calculated as the reciprocal of the absolute risk reduction (ARR), defined as the difference in cumulative lung cancer incidence between treatment arms over the trial follow-up period. Confidence intervals for the NNT were derived by inverting the corresponding confidence intervals for the ARR.

Mouse Plasma Proteomics Analysis

Blood plasma samples were collected at baseline, 3-, 10- and 15- week timepoints from recombined *EGFR*-dox and wild-type C57BL/6J control mice, exposed to PBS or PM following the established 3-week dosing regimen (all mice were fed a doxycycline-containing diet throughout the experiment). Baseline samples were collected after starting the doxycycline diet but prior to any PBS/PM treatment. Plasma samples were sent to SomaLogic (Boulder, Colorado) and assayed in one batch on the SomaScan® 11K v5.0 platform, with samples split evenly between two plates.

Four technical replicates were sent blinded to SomaScan® to assess technical variability. Proteins exhibiting a coefficient of variation (CV) exceeding the 95th percentile within the four technical replicates (pooled mouse plasma with mixed age and sex) were excluded from further analysis, resulting in the removal of 539 proteins. Data were then log-transformed to stabilise variance and approximate a normal distribution, facilitating downstream statistical analyses. One mouse was excluded from analysis as an outlier identified by principal component analysis. Subsequently, for each mouse, changes in the relative fluorescence unit (RFU) at time points 3, 10 and 15 were calculated relative to their respective baseline measurements by subtracting the baseline aptamer RFU from each subsequent time point measurement. A linear mixed-effects model for the 14 proteins within each condition was then fitted using the lme4 (v1.1-37) package using the equation Baseline RFU $\sim$ Time + (1 | Individual Mouse), where Baseline RFU is the baseline value for that mouse. Estimated marginal means were computed within each condition and time point using the emmeans (v1.11.1) package. All p-values across conditions from these contrasts (time 0 vs 3 weeks, time 0 vs 10 weeks, time 0 vs 15 weeks) were adjusted using the Benjamini–Hochberg procedure with adjusted p-values < 0.05 considered significant.

RT-qPCR of IL-1β-treated foetal organoids

Organoids were seeded into four 24-well plate wells, with two wells maintained as untreated controls (AT2 medium only) and two wells treated with recombinant human interleukin-1 beta (IL-1β; 100 ng/mL, Merck, IL038) for 48 hours. Two independent experiments were performed. Total RNA was extracted using the RNeasy Mini Kit (Qiagen) with 100 μL RLT Plus buffer, further steps were performed according to the manufacturer's protocol. RNA was treated with DNase I using the TURBO DNA-free Kit (Invitrogen) according to the manufacturer's instructions. RNA concentration was measured using a DeNovix spectrophotometer. RT-qPCR was performed with the Luna Universal One-Step RT-qPCR Kit (NEB) according to the manufacturer's protocol with 5 ng RNA per reaction and 45 cycles on a QuantStudio 3 Real-Time PCR System (Applied Biosystems). Relative expression was calculated by the comparative Ct method, with the treated samples being normalised to the mean value of the two respective controls measured on the same plate. Two-way ANOVA followed by Sidak's multiple comparisons test was performed using GraphPad Prism version (v10.1.1).

QUANTIFICATION AND STATISTICAL ANALYSIS

Unless otherwise stated, all statistical tests were two-sided and p-values were adjusted using the Benjamini–Hochberg correction. Statistical analyses were performed in R (v4.4.3) or Python (v3.13.2), with the exception of single-nuclei/cell RNA-seq analyses (performed in R v4.3.2) and bulk RNA-seq of murine tumor tissue (performed in R v4.2.3). For each experiment, the specific statistical test, exact value of n, what n represents (e.g. number of mice, cells, or participants), and measures of centre and dispersion (mean ± SEM, median with IQR) are detailed in the corresponding figure legend. Where relevant, assumptions of the selected statistical tests (e.g. normality) were assessed by visual inspection of Q-Q plots or by Shapiro–Wilk testing prior to analysis. A p-value < 0.05 was considered statistically significant unless otherwise stated.

ADDITIONAL RESOURCES

This work involves data from the following registered clinical trials:

- CANTOS (Canakinumab Anti-inflammatory Thrombosis Outcome Study): ClinicalTrials.gov NCT01327846; https://clinicaltrials.gov/study/NCT01327846
- TALENT (Taiwan Lung Cancer Screening in Never-Smoker Trial): ClinicalTrials.gov NCT02611570; https://clinicaltrials.gov/study/NCT02611570
- TRACERx (Tracking Cancer Evolution through therapy [Rx]): ClinicalTrials.gov NCT01888601; https://clinicaltrials.gov/study/NCT01888601
- COPA (COPD Originates in Polluted Air): ClinicalTrials.gov NCT 02236039; https://clinicaltrials.gov/study/NCT02236039

Supplemental figures

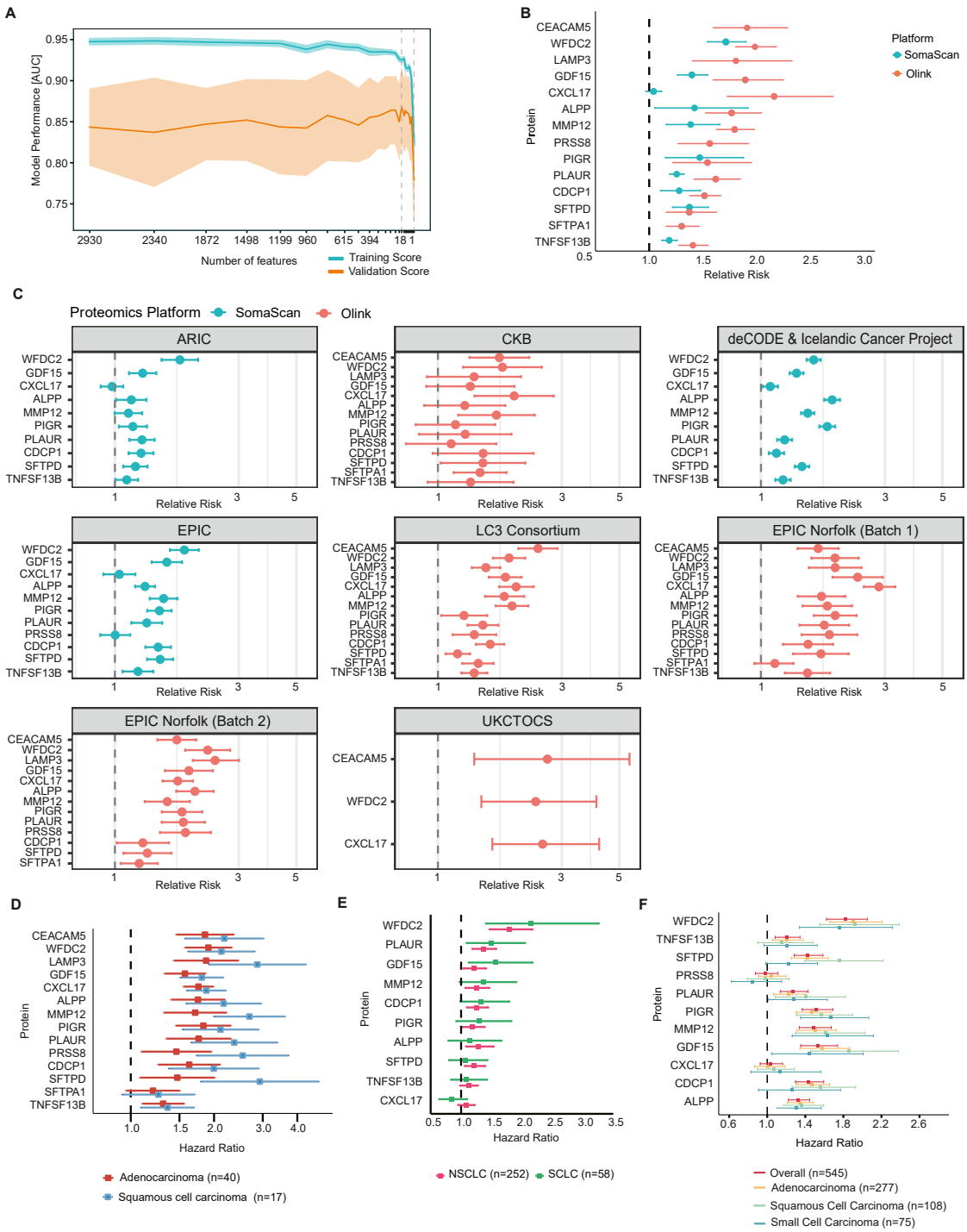


Figure S1. Association of the 14-protein signature with lung cancer in different cohorts, related to Figure 1

(A) Recursive feature elimination on the training set of the UKBB-derived model illustrating change in model performance following feature reduction. Shade represents 95% CI.

(legend continued on next page)

(B) Meta-analysis of each protein across datasets, colored by platform with 95% CI. Platform-specific estimates are displayed only when at least two datasets per platform were available.
(C) Relative risk of each protein with 95% CI for lung cancer in each individual cohort, colored by platform.
(D) Hazard ratio with 95% CI between protein and lung cancer in the UKBB, split between squamous cell carcinoma and adenocarcinoma subtypes.
(E) Hazard ratio with 95% CI between protein and lung cancer in the ARIC cohort, split between non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC).
(F) Hazard ratio with 95% CI between protein and lung cancer in the EPIC cohort, split between adenocarcinoma, squamous cell carcinoma, and small cell carcinoma.

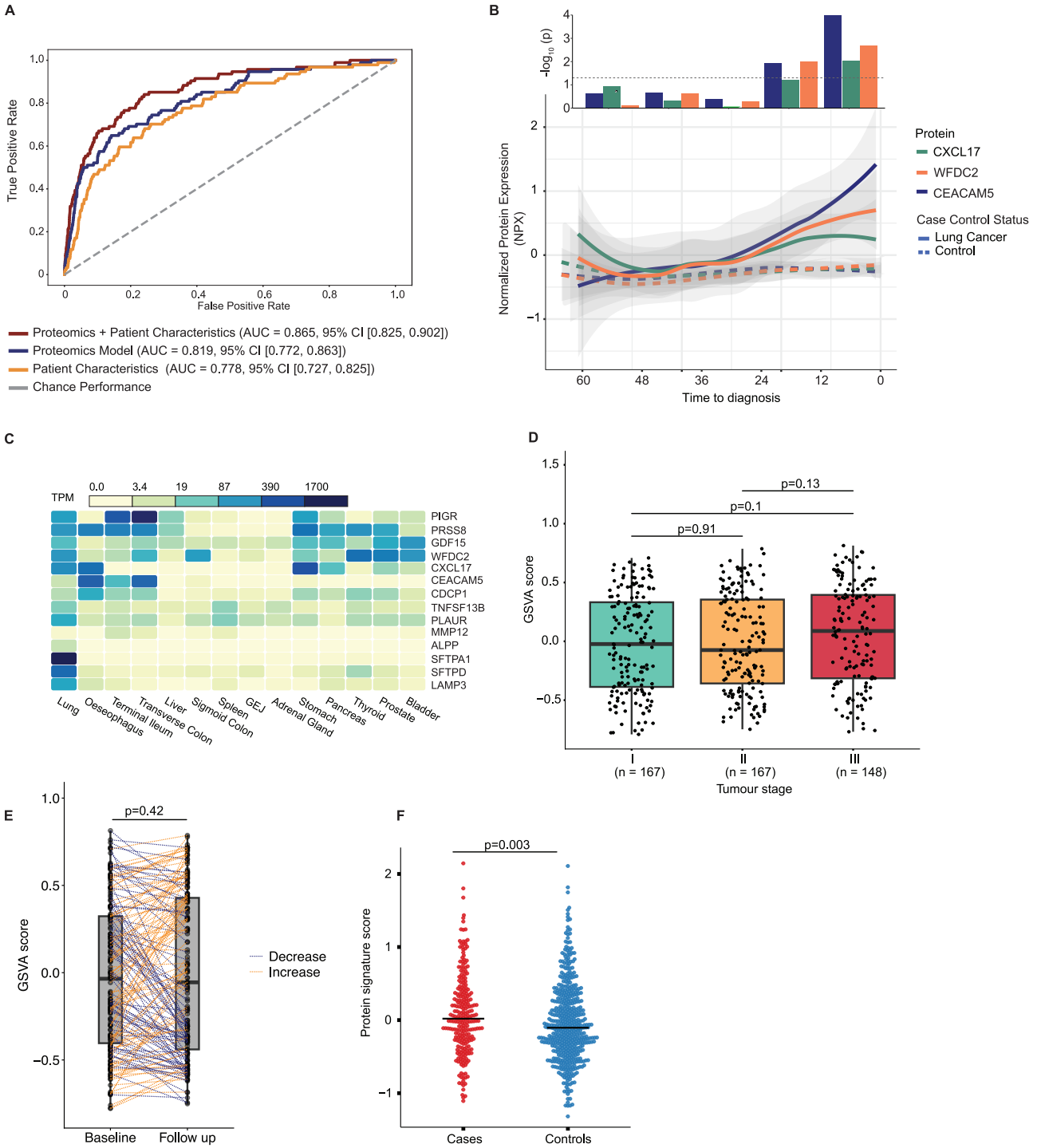


Figure S2. Performance and biological characterization of the 14-protein signature, related to Figure 1

(A) ROC-AUC on the held-out dataset of 12,025 individuals from the UKBB of the proteomics and four patient characteristics model, compared against a model trained on the 14 proteins alone (comparison between the protein-only and clinical-characteristic-only models, $p = 0.26$ by DeLong's test).

(B) LOESS curves of the 3 proteins measured longitudinally at 1-year intervals in 15 never-smokers who were prospectively diagnosed with lung cancer during follow-up and 150 controls from the UKCTOCS clinical trial. Data represent 5 samples per individual before the diagnosis of lung cancer. Shade represents 95% CI. Bars represent the significance of each protein (Wilcoxon test, capped at $-\log_{10} p = 4$).

(C) Expression of each of the 14 proteins of interest across healthy tissues in the GTEx consortium ($n = 19.788$ samples, 946 donors).

(legend continued on next page)

(D) Boxplot of GSVA score with interquartile range of the 14-protein signature in the individuals from the TRACERx observational study across all individuals, by pathological stage at diagnosis (Wilcoxon test).
(E) Paired comparison of the GSVA score of the 14 proteins of interest in individuals from the TRACERx observational study at baseline who did not relapse within 2 years of surgery, with a second sample taken at the last available follow-up (Wilcoxon test).
(F) Aggregate protein signature score (calculated using 10/14 proteins available) between lung cancer cases and controls in the TALENT cohort (Wilcoxon test).

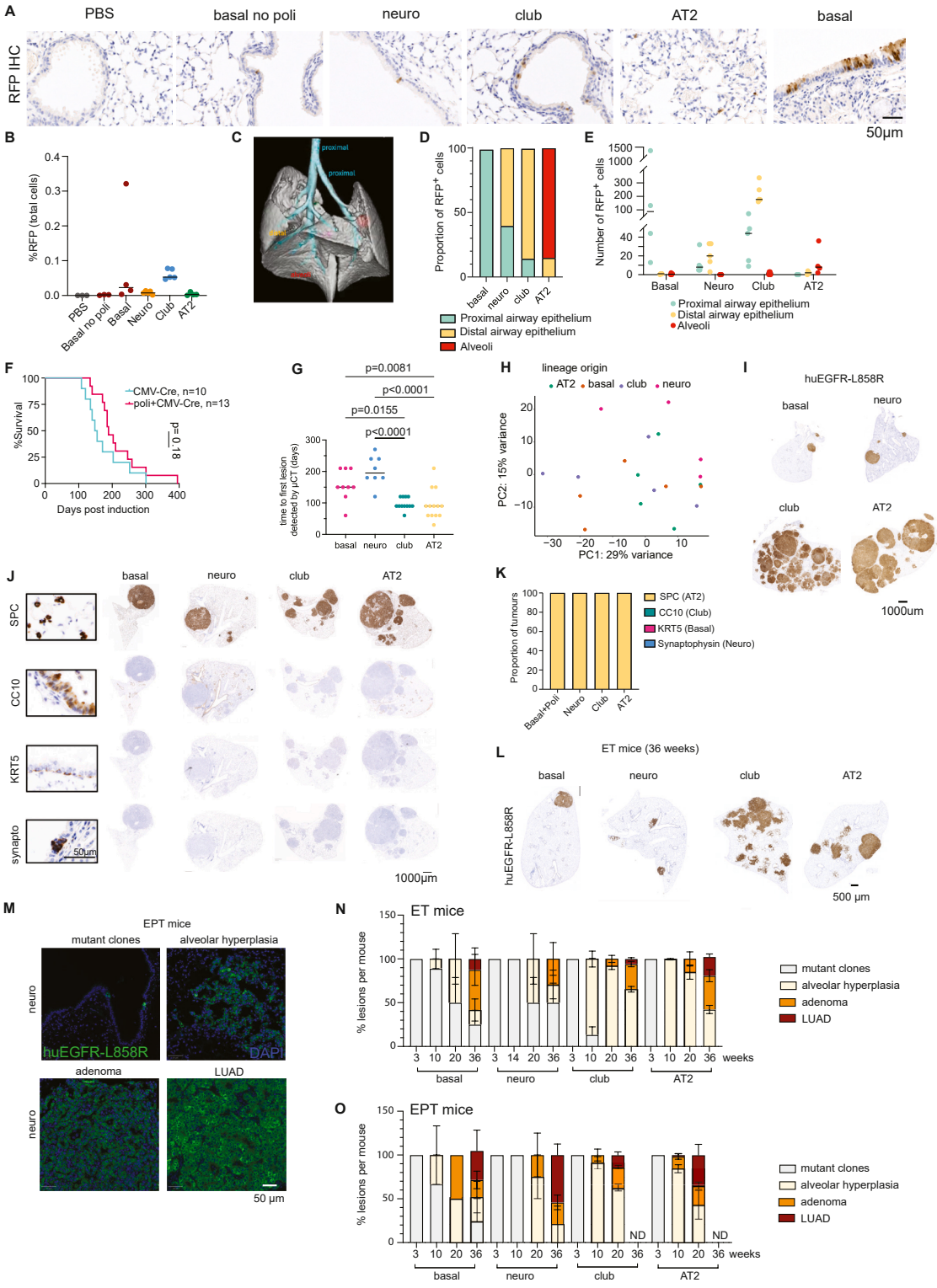


Figure S3. Distinct initiating lineages form *EGFR*-driven LUAD, related to Figure 2

(A and B) Representative images (A) and quantification (B) of tdTomato^{+} cells from immunohistochemistry IHC staining for RFP in the indicated groups in T mice 2 weeks post viral induction. Mice receiving basal cell virus with no polidocanol injury are expected to have limited recombination. The proportion of positive cells is determined relative to total lung cells.

(legend continued on next page)

(C) 3D reconstruction of murine EPT lung from microCT data, demonstrating 3D spatial context of proximal airways (trachea and mainstem bronchi), distal airways (bronchioles), and alveolar tissue (5 lung lobes).
(D and E) Spatial quantification of the proportion (D) and number (E) of RFP^+ cells in T mice in each location. $TdTomato^+$ cells were automatically detected in QuPath and manually assigned to location.
(F) Survival of EPT mice induced with ubiquitous virus (Ad5-CMV-Cre) in control (blue) and polidocanol-treated (red) conditions, collated across two independent cohorts. Mice excluded from survival analysis if collected for non-cancer-related illness, $n = 10$ Ad5-CMV-Cre; $n = 13$ Ad5-CMV-Cre and polidocanol. Mantel-Cox test, $p = 0.18$.
(G) Time to first microCT-detected lesion (> 0.3 mm in diameter) in EPT mice scanned monthly. Data collated across two independent cohorts, and data represent median. $n = 9$ basal, 8 neuroendocrine, 12 club, 13 AT2; One-way ANOVA with Tukey's multiple comparison test.
(H) Principal-component analysis of bulk RNA-seq data generated from tumors collected from EPT mice at ethical endpoint ($n = 5$ tumors from 5 individual mice for basal, club, AT2; $n = 4$ tumors from 4 individual mice for neuroendocrine).
(I) Representative IHC for human EGFR-L858R of EPT lungs induced with lineage-restricted viruses collected at the ethical endpoint. $n = 9$ basal-, 7 neuroendocrine-, 12 club-, and 13 AT2-targeted EPT mice.
(J) Inset: representative adjacent non-malignant lung tissue demonstrating IHC positive control staining of AT2 cells in the alveolar microenvironment (SPC), club cells in bronchioles (CC10), basal cells lining large bronchi (Krt5), and neuroepithelial bodies in bronchioles (synaptophysin). Scale bar is 50 μm. Main panels: representative serial sections of lung lobes from EPT tumors derived from basal, neuroendocrine, club, and AT2 lineages collected at the ethical endpoint. Scale bar is 1,000 μm.
(K) Quantification of IHC data shown in (J). $n = 9$ basal-, 7 neuroendocrine-, 12 club-, and 13 AT2-induced EPT mice analyzed.
(L) Representative IHC for human EGFR-L858R protein in basal, club, and AT2-targeted ET mice, 36 weeks post-oncogene induction. Scale bar is 500 μm in all images.
(M) Representative immunofluorescence of neuroendocrine-targeted EPT mice demonstrating mutant clones (3 weeks post-induction), alveolar hyperplasia (20 weeks), adenoma (20 weeks), and LUAD (36 weeks).
(N and O) Quantification of the proportion of lesion types detected per (N) ET mouse over time. 3 weeks, $n = 3$ basal, neuro, club, and AT2. 10 weeks, $n = 9$ basal, $n = 15$ club, $n = 14$ AT2; 14 weeks, $n = 3$ neuro; 20 weeks, $n = 4$ basal, 4 neuro, $n = 3$ club, and $n = 3$ AT2. 36 weeks; $n = 8$ basal, $n = 4$ neuroendocrine, $n = 11$ club, $n = 8$ AT2-targeted ET mice, and (O) per EPT mouse over time. 3 weeks, $n = 2$ basal, $n = 3$ neuro, club, and AT2. 10 weeks; $n = 3$ basal, neuro, club, and AT2. 20 weeks, $n = 1$ basal, $n = 2$ neuro, $n = 3$ club, $n = 2$ AT2. 36 weeks, $n = 4$ basal, $n = 2$ neuro. ND, not determined due to time point occurring beyond the ethical endpoint of these animals. Time to adenocarcinoma (LUAD) onset in EPT vs. ET mice; AT2: 10 vs. 36 weeks, $p = 1.0 \times 10^{-4}$; club: 20 vs. 36 weeks, $p = 0.002$; basal: 36 vs. 36 weeks, non-significant; neuroendocrine: 36 weeks vs. no LUAD detected in ET mice; Fisher's exact test. Adenocarcinoma onset at 20 weeks in EPT mice across lineage groups: AT2/club 6/6 mice vs. basal/neuroendocrine 0/4 mice; Fisher's exact $p = 0.0048$. Bars represent mean ± SEM.

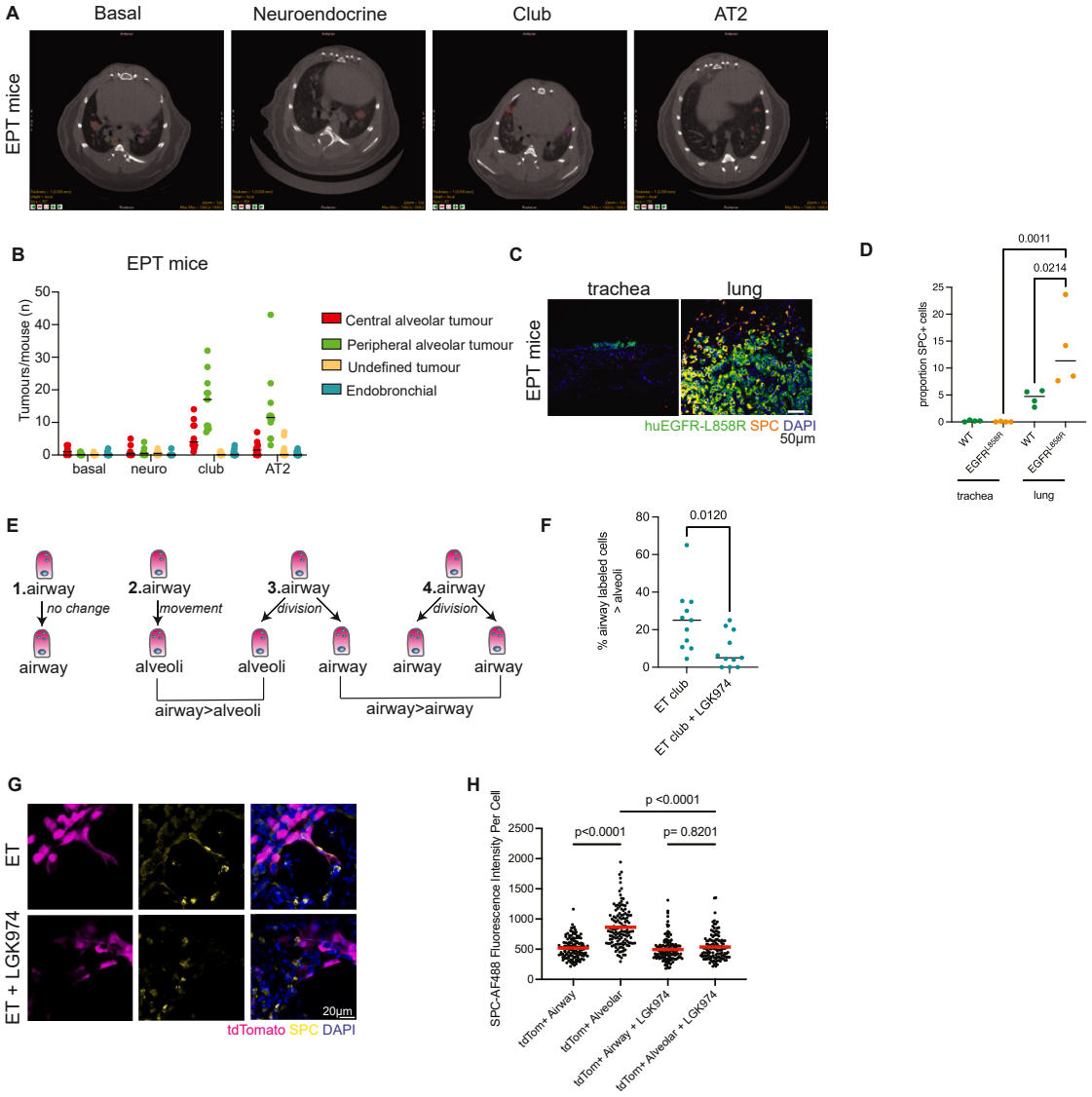


Figure S4. Airway and alveolar initiating cells form lesions in the alveolar niche, related to Figure 2

(A) Representative axial views of microCT data of first-detected lesions in EPT mice induced with each virus. *n* = 9 basal-, 8 neuroendocrine-, 12 club-, and 13 AT2-targeted EPT mice analyzed. Lesions are outlined by colored lines; for full experimental results, see Videos S1, S2, S3, and S4.

(B) Number of tumors in EPT mice at ethical endpoint, separated by location within the lung. Central alveolar tumor: occurring near proximal airways in the alveolar compartment; distal alveolar tumor: occurring at lung periphery in the alveolar compartment; endobronchial tumor: tumor within the airway lumen. Undefined: tumor occupying entire lung lobe (location indeterminate).

(C) Representative image of dual human EGFR-L858R and SPC immunofluorescence of *EGFR*-mutant basal-derived cells in the trachea (left) and lung (right) of an EPT mouse collected 7 months post-induction with polidocanol and the basal cell virus.

(D) Quantification of data in (C) showing WT or *EGFR*-mutant cells that are SPC+ in the trachea and lung of basal-induced EPT mice collected at ethical endpoint (7–12 months post-oncogene induction), *n* = 4. One-way ANOVA with Sidak's multiple comparison test.

(E) Schematic of quantification of "airway > alveoli" and "airway > airway" phenotype observed in PCLS experiment from T and ET mice targeted with club cell virus, depicting the cell division and cell movement that underlies this classification.

(F) PCLS from club cell-targeted ET mice treated with DMSO or 100 nM LGK974 for 72 hours *ex vivo*, derived 14 weeks (repeat 1, 2 animals) or 20 weeks (repeat 2, 2 animals) post-induction. One representative experiment shown; *n* = 2 mice per condition, 5–6 fields of view per animal; unpaired *t* test.

(G and H) Serial-sectioned PCLS from the same experiment as (F), separate PCLS simultaneously cultured and stained for SPC-AF488. (G) Representative images. (H) Quantification of tdTomato+ SPC-expressing cells in airway and alveolar compartments ± LGK974 (Kruskal-Wallis test); one representative experiment of two independent experiments, *n* = 2 mice per condition.

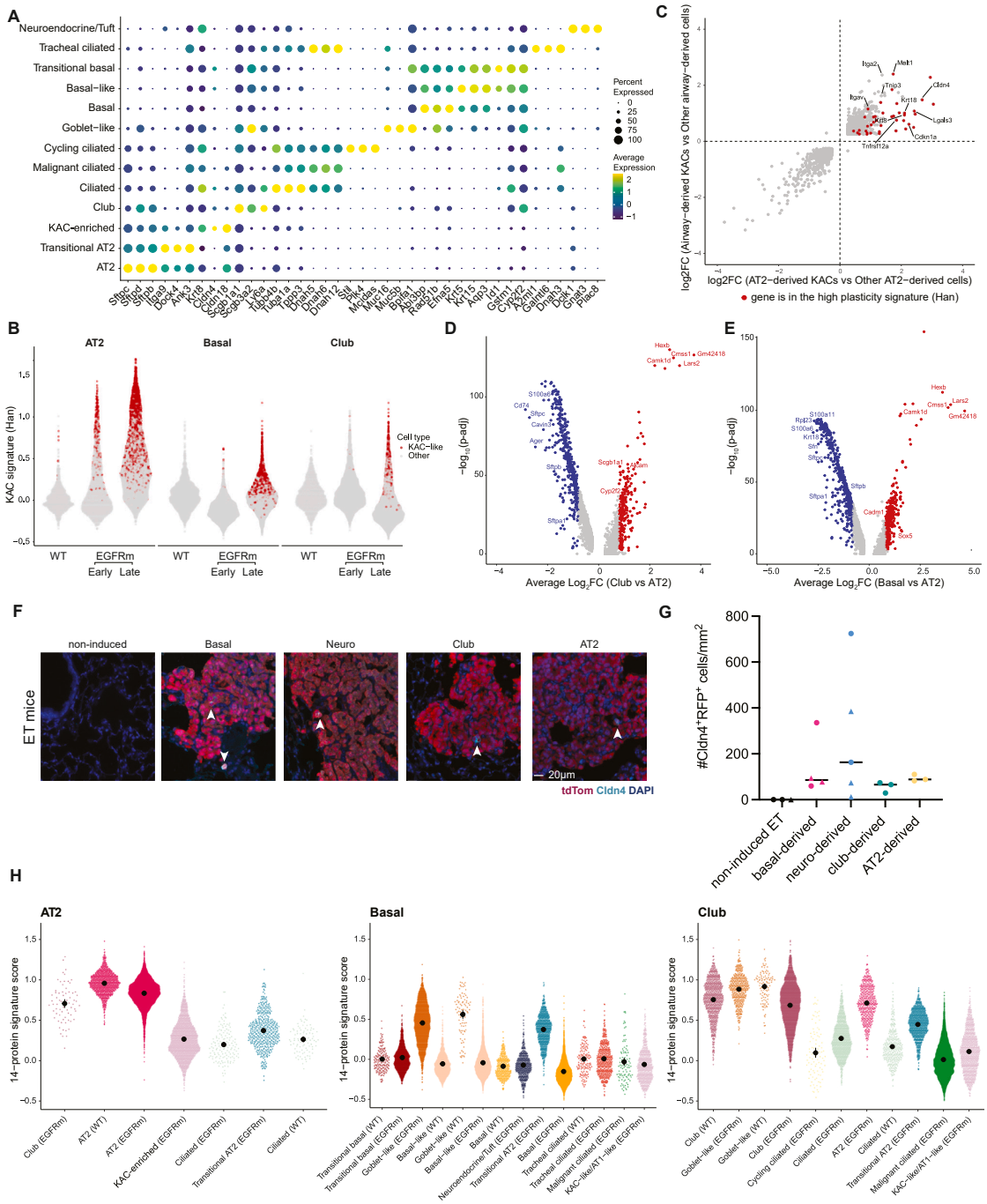


Figure S5. Lung epithelial lineages converge on alveolar intermediate cell states upon acquiring an *EGFR* mutation, related to Figure 2

(A) Bubble plot depicting the expression of marker genes of cellular subsets underlying cell classification in Figure 2I.

(B) Violin plot showing the KAC gene signature score expression (from Han et al.[17]) for AT2-, basal-, and club-derived cells, showing all cells detected in each condition per time point. KACs derived from each lineage, as identified from cell clustering in Figure 2I, are highlighted in red.

(C) Comparison of genes enriched in airway (basal and club)-derived KACs compared with other airway cells, and genes enriched in AT2-derived KACs compared with other alveolar cells, demonstrating common markers of KACs; absolute $\log_2$FC > 0.1, p value < 0.05. Genes are colored in red if also present in the Han et al. mouse KAC's signature.[17]

(D and E) Volcano plots depicting significantly upregulated (red) and downregulated (blue) genes in basal cell-derived KACs (D) or club cell-derived KACs (E) compared with AT2-derived KACs. Differential gene expression determined by absolute $\log_2$FC $\geq$ 0.8 and adjusted p value < 0.01.

(F and G) Representative images (F) and quantification (G) of CLDN4$^+$ RFP$^+$ cells (indicated with white arrows) within RFP$^+$ hyperplasias and adenomas in the alveolar compartment from ET mice collected 20 (depicted in circles) and 36 weeks (depicted in triangles) post-oncogene induction, with non-induced ET mice as

(legend continued on next page)

a control. Data represent the average number of CLDN4^{+} EGFR-L858R^{+} per mm^{2} observed across multiple lesions per mouse. No RFP^{+} lesions were observed in airways. *n* = 3 AT2 (20 weeks), *n* = 3 club (20 weeks), *n* = 5 neuro (20 weeks and 36 weeks), and *n* = 4 basal (20 weeks and 36 weeks). One-way ANOVA, ns. (H) Quantifying the 14-protein gene set signature score within AT2-, basal, or club-lineage- tagged cells in ET tumorigenesis, with cell clusters identified within each lineage (determined from total cell clustering in Figure 2I) ordered along the *x* axis by median pseudo-time; black dot represents group mean signature score.

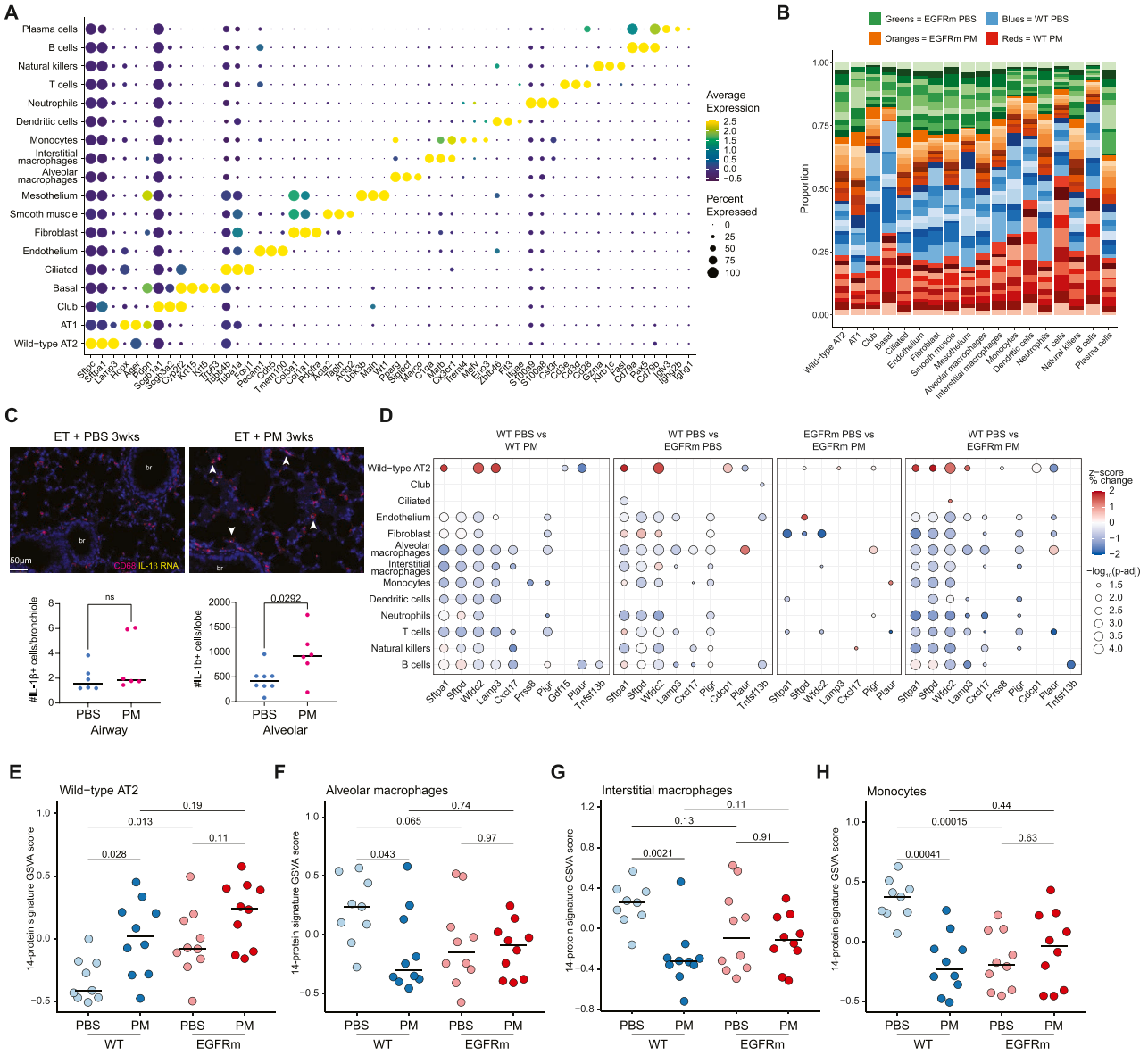


Figure S6. Tumor promotion remodels the lung microenvironment, related to Figure 3

(A) Dot plot of cluster-specific genes in the mouse lung microenvironment scRNA-seq dataset.

(B) Stacked bar chart showing the distribution of cells per mouse (from hashing) per cell type. Wild-type (WT) PM, red shades; WT PBS, blue shades; *EGFR*-mutant PBS, green shades; *EGFR*-mutant PM, orange shades.

(C) Spatial quantification of IL-1β RNAscope and CD68 (macrophage) of ET mice 3 weeks post PM exposure in the alveolar or airway compartment. Peri-airway regions are defined as 50 microns from bronchioles (br), and arrows indicate double-positive cells. $n = 3$ mice, 2–3 lobes per mouse; bronchiole data are averaged from 15 to 40 airways/lobe. Unpaired *t* test.

(D) Bubble plot comparing expression of the 12 detected protein transcripts between conditions in all cell types (where *n* cells > 100) within the dataset. Color of the bubble denotes the per-gene normalized percentage change; size represents $-\log_{10} p$ value (Wilcoxon test), where only results with $p < 0.05$ are shown.

(E–H) GSVA score of the 14-protein signature on pseudobulked WT AT2 cells (E), alveolar macrophages (F), interstitial macrophages (G), and monocytes (H). Wilcoxon test.

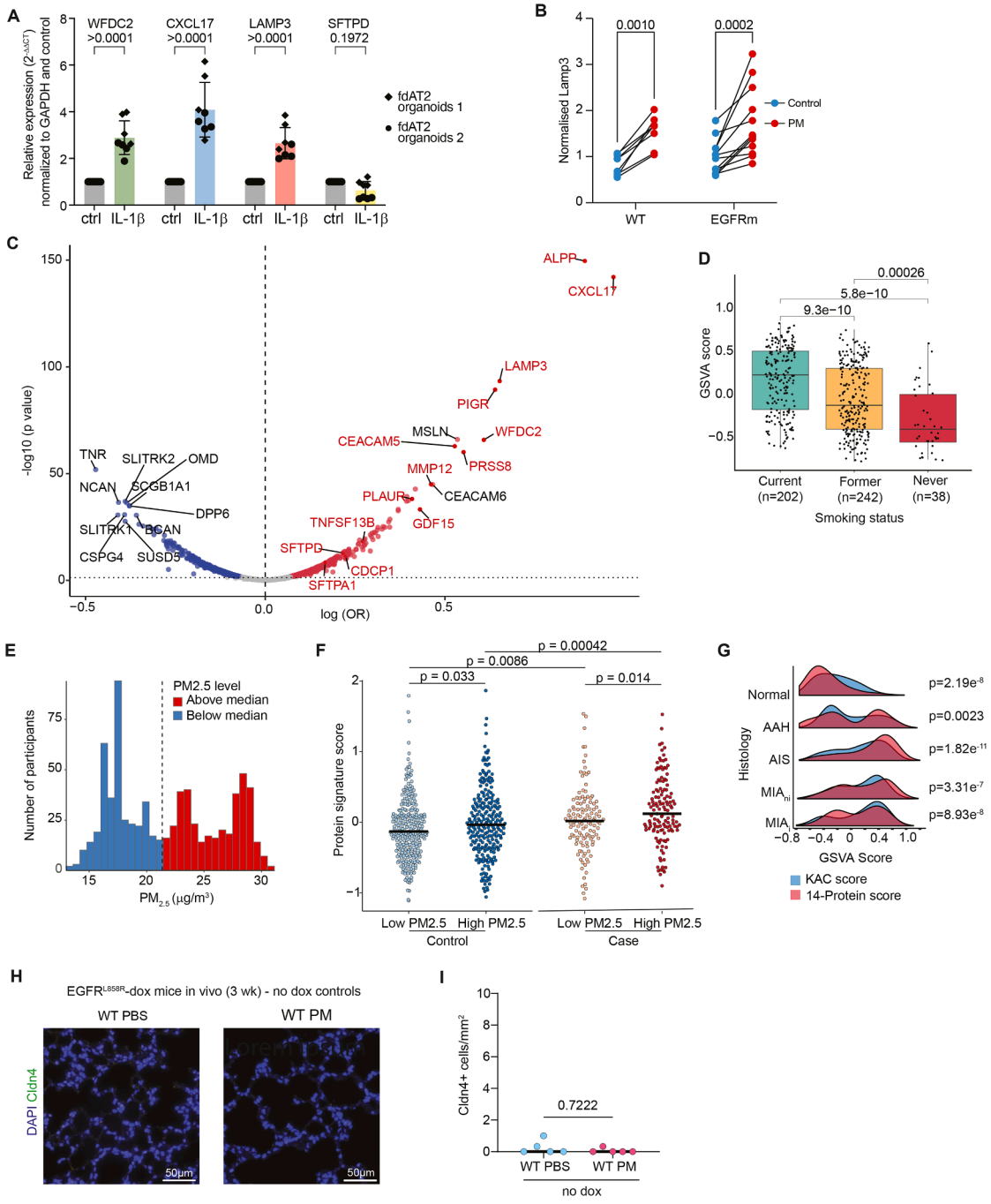


Figure S7. Tumor-promoting inflammation regulates expression of the 14-protein signature, related to Figures 3 and 4

(A) RT-PCR from 48 hours of IL-1β treatment of human WT fetal-lung-derived AT2 (fdAT2) organoids of four epithelial proteins compared with control (two-way ANOVA) between conditions (n = 2 separate samples).

(B) ELISA assay of LAMP3 protein levels in supernatants of PCLS generated from *EGFR*-dox mice either untreated (WT) or treated with doxycycline for 5 days *in vivo* before *ex vivo* challenge with 50 μg/mL PM for 72 h. Data are normalized to PBS control, and paired t tests between PCLS from the same mouse.

(C) Volcano plot from the UKBB held-out set comparing the relationship between ever-smokers (current or previous smokers) and never-smokers after adjusting for age, sex, BMI, and incident diagnosis of lung cancer, COPD, or IPF. Top 10 significantly up- and downregulated proteins labeled alongside 14 proteins of interest (in red). Gray line represents proteins significantly associated by Wilcoxon test.

(D) Boxplot of GSVA scores with intequartile ranges of the 14 proteins of interest for baseline samples from individuals in the TRACERx cohort, split by smoking status.

(E) Distribution of $PM_{2.5}$ levels among individuals in the TALENT cohort.

(legend continued on next page)

(F) Aggregated protein signature score in individuals who developed lung cancer and controls, split by median $PM_{2.5}$ levels. Significance is determined by Benjamini-Hochberg-adjusted *p* values (Wilcoxon test).
(G) Ridge plot between the 14-protein signature and the KAC signature highlighting overlap across tissues ($p < 0.05$ in all tissues using linear regression).
(H and I) (H) Visualization and (I) quantification of claudin 4 positive ($Cldn4^+$) cells in WT mice (*EGFR*-dox mice not treated with doxycycline) exposed to PM and harvested directly after the last PM exposure, $n = 5$.

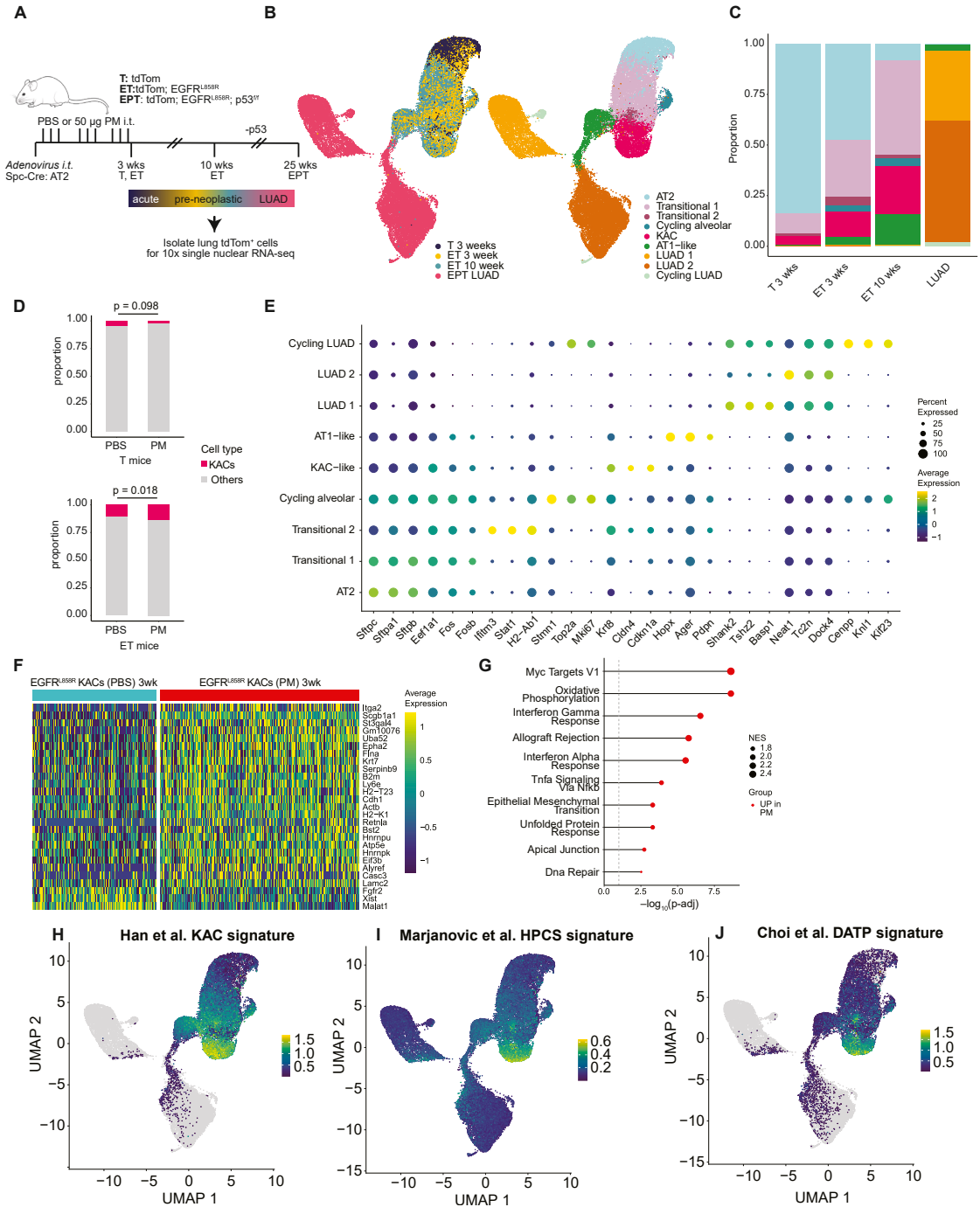


Figure S8. PM exposure expands and reprograms *EGFR*-mutant KACs, related to Figure 4

(A) Experimental schematic depicting T, ET, and EPT mouse models of *EGFR*-driven LUAD, induced with Ad5-SPC-Cre via intratracheal instillation (i.t.), following PBS control or PM exposure. Lungs were harvested acutely post-exposure (3 weeks, 10 mice per treatment), or at 10 weeks ($n = 10$ mice per treatment). EPT mice ($n = 2$) were collected at the ethical endpoint, approximately 25 weeks post induction. At each time point, tdTomato$^+$ EpCAM$^+$ cells were fluorescence-activated cell sorting (FACS)-isolated and subject to snRNA-seq to profile AT2-lineage-tagged cells. PBS control data correspond to the SPC-Cre-induced condition analyzed earlier as part of cell-of-origin analyses described in Figure 2H.

(B) UMAP visualization of lung tumorigenesis profiled by snRNA-seq, labelled by condition (left) and cell state (right).

(C) Proportion quantification of cell states across conditions.

(D) Proportion quantification of KACs from PBS- and PM-exposed control and *EGFR*-mutant conditions at 3 weeks. Chi-squared test.

(E) Dot plot of cluster-specific markers from the mouse lung tumorigenesis snRNA-seq dataset.

(legend continued on next page)

(F) Heatmap of differentially expressed genes comparing PBS- and PM-exposed *EGFR*-mutant KACs at 3 weeks. Cap of ±1.2 applied to scale.
(G) Differentially upregulated Hallmark gene sets in KACs from PM- and PBS-exposed conditions at 3 weeks.
(H–J) UMAP visualization of lung tumorigenesis profiled by snRNA-seq as described above with (H) projection of the Han et al. mouse KAC signature,[17] (I) Marjanovic et al. HPCS signature,[16] and (J) Choi et al. DATP signature.[21]

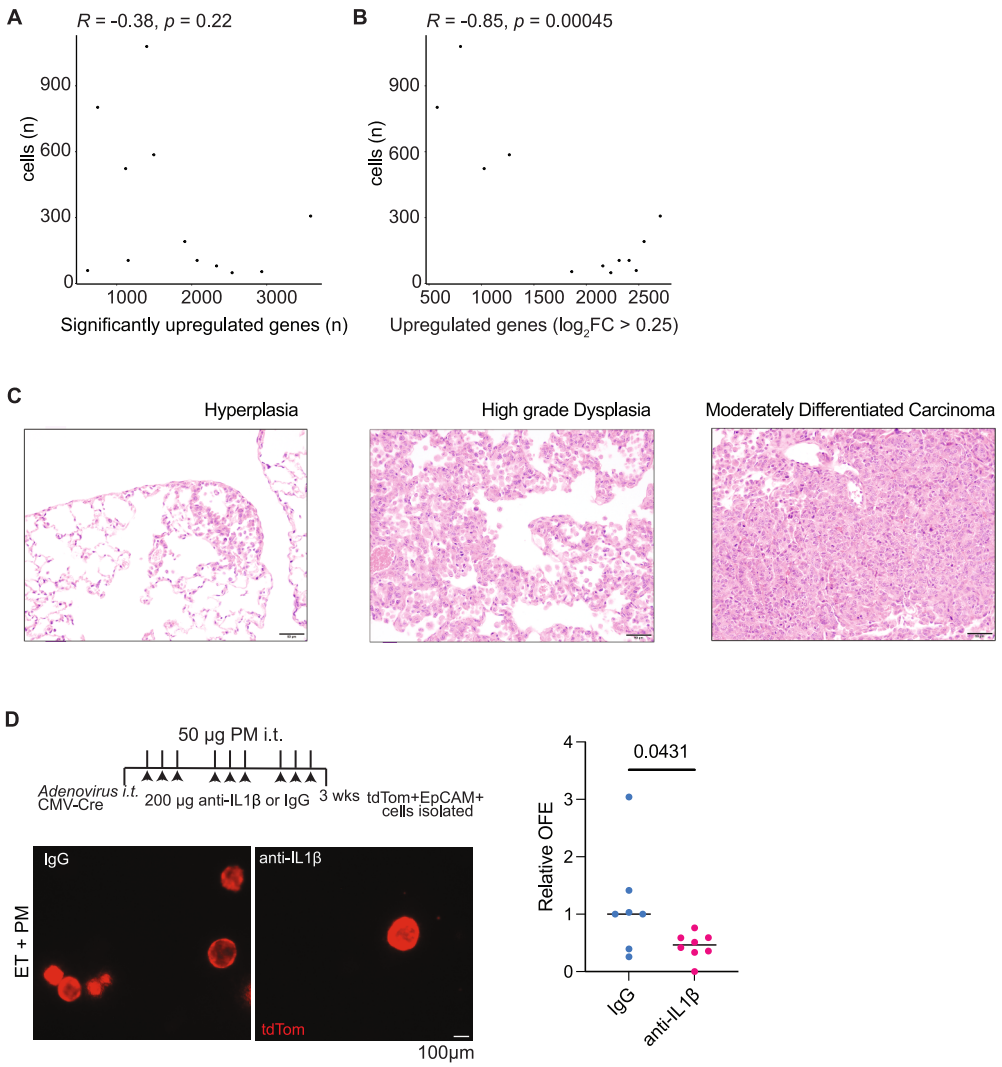


Figure S9. IL-1β blockade administered concurrently with PM exposure reduces the organoid-forming efficiency of *EGFR*-mutant epithelial cells, related to Figure 4

(A and B) Correlation plots depicting the relationship between the number of cells and the number of significantly differentially expressed genes (A) and the number of differentially expressed genes where $\log_2$ fold change ($\log_2$FC) > 0.25 (B) (Pearson's correlation).

(C) Representative images of hyperplasia, high-grade dysplasia, and moderately differentiated carcinoma in relation to Figure 4K. Scale bar, 50 µm.

(D) Representative images and quantification of organoid-forming efficiency of tdTomato$^+$ *EGFR*-mutant epithelial cells isolated at 3 weeks post-oncogene induction, from anti-IL-1β (200 µg) or hamster IgG control-treated ET mice, administered *in vivo* concurrently with PM exposure (50 µg), 3 exposures per week for 3 weeks. Data compiled from $n = 3$ independent experiments, $n = 7$ IgG controls, and $n = 8$ anti-IL-1β-treated mice.