Nature. Author manuscript; available in PMC 2023 June 01.

Published in final edited form as:
Nature. 2023 April 01; 616(7955): 159–167. doi:10.1038/s41586-023-05874-3.

[128]Corresponding author: Charles.Swanton@crick.ac.uk.
[127]These authors jointly supervised this work: Emilia L. Lim, James DeGregori, Mariam Jamal-Hanjani.
*A list of authors and their affiliations appears at the end of the paper.

Author contributions

W.H. and E.L.L jointly conceptualized the project, designed the project, performed the experiments analyses and wrote the manuscript. W.H. led and performed the mouse experiments, and E.L.L. led and performed the bioinformatics and epidemiology analyses. C.E.W. performed the mouse experiments, helped write the manuscript and curated the mutation literature. C.L. performed the human RNA-seq analyses and curated the pollution data. M.A. performed the UKBB analyses. K.C. assembled and analysed the TRACERx cohort dataset. F.-C.K. and M.-H.L. performed the Taiwan epidemiological analyses. F. Marongiu, E.J.E.J., C.I.T., M.G., YE.M., D.T.M. and R.L.K. generated and analysed the Duplex-seq data. O.P. wrote the Duplex-seq bioinformatics pipeline and performed the mutational signature analyses. H.C. and S.-H.L. performed the Korea epidemiological analyses. F.v.M., J.B., A.M. and D.R.C. were involved in mouse data acquisition. F.S.R. was involved with organoid experiments. S.V., A.R. and C.N.-L. curated and performed DNA extractions on TRACERx and PEACE samples. T.K. helped analyse patient clinical characteristics. D.M., S.D. and M.S. performed pathological assessments of human tissue samples. A.N., B.B., J.R.M.B. and C.M.-R. performed mouse RNA-seq analyses. M.H.R., R.D.H. and S.L. designed and generated data for the human crossover study. A.S.-B. And S.L.P. were involved in mouse pathology analyses. M.L., K. Lavelle, J.P., S.H. and F.E.M. curated England's National Disease Registration Service data et. R.M. curated the Canadian cohort. M.A.B. and C.B. wrote and ran the MiSeq pipeline. C.A., L.H.S., Y.C. and A.M.G. performed the ddPCR experiments. I.M., J. Downward, T.J., N.K. and E.G. provided supervision for the mouse experiments. M.-J.F., M.H., P.A. and N.M. provided guidance and supervision for the bioinformatics analyses. S.L., P.S., R.H., C.T., C. D.B., A.H. and K. Litchfield provided supervision for epidemiological analyses. C.C. provided supervision for the human crossover study. J. DeGregori designed the BDRE study and supervised the healthy tissue profiling work. M.J.-H. designed the PEACE and TRACERx study protocols, and E.L.L. and M.J.-H. jointly supervised the study and collaborations. C.S. supervised the work, provided strategic oversight and helped write the manuscript.

Competing interests

M.A.B. has consulted for Achilles Therapeutics. L.H.S., Y.C. and A.M.G. have ownership interest in SAGA Diagnostics. S.V. is a co-inventor to a patent to detecting molecules in a sample (US patent 10578620). D.A.M. reports speaker fees from AstraZeneca, Eli Lilly and Takeda, consultancy fees from AstraZeneca, Thermo Fisher, Takeda, Amgen, Janssen, MIM Software, Bristol-Myers Squibb and Eli Lilly, and has received educational support from Takeda and Amgen. C.A. has received speaking honoraria or expenses from Novartis, Roche, AstraZeneca and Bristol-Myers Squibb and reports employment at AstraZeneca. C.A. is an inventor on a European patent application relating to assay technology to detect tumour recurrence (PCT/GB2017/053289). The patent has been licensed to commercial entities and under their terms of employment, C.A is due a revenue share of any revenue generated from such licence(s). C.A. declares a patent application (PCT/US2017/028013) for methods to detect lung cancer. C.A. is a named inventor on a patent application to determine methods and systems for tumour monitoring (PCT/EP2022/077987). T.J. is a member of the Board of Directors of Amgen and Thermo Fisher Scientific, and a co-Founder of Dragonfly Therapeutics and T2 Biosystems. T.J. serves on the Scientific Advisory Board (SAB) of Dragonfly Therapeutics, SQZ Biotech and Skyhawk Therapeutics. T.J. is also President of Break Through Cancer. K. Litchfield has a patent on indel burden and CPI response pending and speaker fees from Roche tissue diagnostics, research funding from CRUK TDL–Ono–LifeArc alliance, Genesis Therapeutics, and consulting roles with Ellipses Pharma, Monopteros and Kynos Therapeutics. N.M. has received consultancy fees and has stock options in Achilles Therapeutics. N.M. holds European patents relating to targeting neoantigens (PCT/EP2016/059401), identifying patient response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004), and predicting survival rates of patients with cancer (PCT/GB2020/050221). C.T. has received honoraria for educational activities and advisory boards from AstraZeneca and Roche (all proceeds donated to registered charity 11511580). C.D.B. has consultantships with GRAIL, LLC, NHS Galleri Trial, IDMC, Mercy BioAnalytics, Lucid DX and Medial EarlySign. J. Downward has acted as a consultant for AstraZeneca, Jubilant, Theras, Roche and Vividion and has funded research agreements with Bristol-Myers Squibb, Revolution Medicines and AstraZeneca. A.H. has received fees for being a member of Independent Data Monitoring Committees for Roche-sponsored clinical trials, and academic projects co-ordinated by Roche. M.J.-H. is a CRUK Career Establishment Awardee and has received funding from CRUK, NIH National Cancer Institute, IASLC International Lung Cancer Foundation, Lung Cancer Research Foundation, Rosetrees Trust, UKI NETs, NIHR, NIHR UCLH Biomedical Research Centre. M.J.-H. has consulted for, and is a member of, the Achilles Therapeutics Scientific Advisory Board and Steering Committee, has received speaker honoraria from Pfizer, Astex Pharmaceuticals, Oslo Cancer Cluster, Bristol Myers Squibb, and is co-inventor on a European patent application relating to methods to detect lung cancer PCT/US2017/028013).M.G., Y.E.M., R.L.K. and D.T.M. acknowledge grant support from Bristol-Myers Squibb. C.S. acknowledges grant support from AstraZeneca, Boehringer-Ingelheim, Bristol-Myers Squibb, Pfizer, Roche-Ventana, Invitae (previously Archer Dx Inc-collaboration in minimal residual disease sequencing technologies), and Ono Pharmaceutical. C.S. is an AstraZeneca Advisory Board member and Chief Investigator for the AZ MeRmaiD 1 and 2 clinical trials and is also Co-Chief Investigator of the NHS Galleri trial funded by GRAIL and a paid member of GRAIL's SAB. He receives consultant fees from Achilles Therapeutics (also SAB member), Bicycle Therapeutics (also a SAB member), Genentech, Medicxi, Roche Innovation Centre–Shanghai, Metabomed (until July 2022), and the Sarah Cannon Research Institute. He had stock options in Apogen Biotechnologies and GRAIL until June 2021, and currently has stock options in Epic Bioscience, Bicycle Therapeutics, and has stock options and is co-founder of Achilles Therapeutics. C.S. is an inventor on a European patent application relating to assay technology to detect tumour recurrence (PCT/GB2017/053289), the patent has been licensed to commercial entities, and under his terms of

Europe PMC Funders Author Manuscripts

Lung adenocarcinoma promotion by air pollutants

A full list of authors and affiliations appears at the end of the article.

[#] These authors contributed equally to this work.

Abstract

A complete understanding of how exposure to environmental substances promotes cancer formation is lacking. More than 70 years ago, tumorigenesis was proposed to occur in a two-step process: an initiating step that induces mutations in healthy cells, followed by a promoter step that triggers cancer development[1]. Here we propose that environmental particulate matter measuring ≤2.5 μm ($PM_{2.5}$), known to be associated with lung cancer risk, promotes lung cancer by acting on cells that harbour pre-existing oncogenic mutations in healthy lung tissue. Focusing on EGFR-driven lung cancer, which is more common in never-smokers or light smokers, we found a significant association between $PM_{2.5}$ levels and the incidence of lung cancer for 32,957 EGFR driven lung cancer cases in four within-country cohorts. Functional mouse models revealed that air pollutants cause an influx of macrophages into the lung and release of interleukin-1β. This process results in a progenitor-like cell state within EGFR mutant lung alveolar type II epithelial cells that fuels tumorigenesis. Ultradeep mutational profiling of histologically normal lung tissue from 295 individuals across 3 clinical cohorts revealed oncogenic EGFR and KRAS driver mutations in 18% and 53% of healthy tissue samples, respectively. These findings collectively support a tumour promoting role for $PM_{2.5}$ air pollutants and provide impetus for public health policy initiatives to address air pollution to reduce disease burden.

Barrier organs such as the lung are directly affected by exposure to environmental challenges. Accordingly, more than 20 environmental and occupational agents are lung carcinogens[2], and exposure to these are of particular relevance in understanding lung cancer in the never-smoking population. Lung cancer in never-smokers (LCINS) is the eighth most common cause of cancer death in the UK and has distinct clinical and molecular characteristics compared with lung cancer in smokers[3]. LCINS frequently harbour adenocarcinomas with oncogenic EGFR mutations and are more commonly observed in female individuals and in individuals with East Asian ancestry compared with patients with Western ancestry[4]. Several factors have been proposed to explain the observed sex and geographical disparities of lung cancer driven by EGFR mutations, including germline genetics[5], ethnicity, radon exposure, occupational carcinogen exposure and air pollution[6]. Air pollution accounts for 7 million deaths per year, with 99% of people living in areas that exceed World Health Organization guidelines (<5 μg m^{-3} annually)[7]. Particulate

employment, C.S. is due a revenue share of any revenue generated from such licence(s). C.S. holds patents relating to targeting neoantigens (PCT/EP2016/059401), identifying patient response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221), identifying patients who respond to cancer treatment (PCT/GB2018/051912), a US patent relating to detecting tumour mutations (PCT/US2017/28013), methods for lung cancer detection (US20190106751A1) and both a European and US patent related to identifying insertion/deletion mutation targets (PCT/GB2018/051892) and is co-inventor to a patent application to determine methods and systems for tumour monitoring (PCT/EP2022/077987). C.S has received honoraria from Amgen, AstraZeneca, Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol Myers Squibb, Illumina, and Roche-Ventana.

matter (PM) is a key constituent of air pollution and is classified by aerodynamic size. Fine particles ≤2.5 μm ($PM_{2.5}$) are able to travel deep into the lung and are linked to multiple adverse health effects, including heart disease and lung cancer[7]. Traditionally, it is thought that carcinogens cause tumours by directly inducing DNA damage. However, recent data suggest that many carcinogens do not cause a detectable DNA mutational signature in tumours following exposure[8,9]. Genetic analyses of oesophageal cancer showed that mutational signatures do not fully explain the varied geographical incidence of this cancer[10], and efforts that have profiled tumour genomes in LCINS failed to detect a dominant carcinogenic signal of mutations deriving from exogenous sources[11–13]. We propose that air pollutants might promote inflammatory changes in the lung tissue microenvironment that permit pre-existing mutated clones to expand, consistent with the two-stage carcinogenesis model of initiation and promotion[1]. To address this hypothesis, we combined epidemiological evidence with functional preclinical models and clinical cohorts to decipher potential mechanisms of air-pollution-induced lung tumour promotion and actionable targets for molecular cancer prevention (Extended Data Fig. 1a).

Lung cancer incidence and $PM_{2.5}$ levels

In a companion article[14], our analysis of lung adenocarcinoma (LUAD) tumours from the TRACERx 421 cohort revealed that despite a history of smoking, a minority of patients (8%) lacked evidence of smoking-mediated mutagenesis, including 6.4% of patients with >15 years of smoking. Consistent with that analysis, in the current study, 7–12% of smokers in the TRACERx 421 cohort did not have a driver single nucleotide variant that could be attributed to a smoking-related single base substitution (SBS) mutation signature (SBS4 or SBS92) (Extended Data Fig. 1b). This result suggests that smoking may promote cancer through additional mechanisms[15]. To understand whether air pollutants can promote the formation of lung tumours without inducing exogenous mutational signatures, we studied EGFR-driven lung cancer, which has a high prevalence in LCINS (in England, the probability that a LCINS is caused by an EGFR-driven tumour is 36–40%), owing to its low mutational burden[11–13] and greater incidence in countries in Asia[4] (Supplementary Tables 1–3). To examine the relationship between air pollutants and EGFR-driven lung cancer incidence, we used several ecological correlation analyses, acknowledging that these analyses only provide estimates of incidence. We considered data from three countries to explore different ranges of $PM_{2.5}$ air pollution and ethnicities: England (92% white; $PM_{2.5}$ interquartile range (IQR): 9.95–11.2 μg m^{-3}); South Korea (>99% Asian[16]; $PM_{2.5}$ IQR: 24.0–27.0 μg m^{-3}); and Taiwan (>98% Asian[17]; $PM_{2.5}$ IQR: 24.3–38.2 μg m^{-3}) (Supplemetary Tables 1–3). In each country, there was a consistent relationship between $PM_{2.5}$ levels (average concentration per geographical area) and estimated EGFR-driven lung cancer incidence (Fig. 1a–c). The relative rates of EGFR-driven lung cancer incidence (per 100,000 population), per 1 μg m^{-3} increment of $PM_{2.5}$ levels were 0.63 ($P = 0.0028$) in England, 0.71 ($P = 0.0091$) in South Korea and 1.82 ($P = 4.01 \times 10^{-6}$) in Taiwan. When restricting the English cohort to adenocarcinoma cases, the relationship remained significant (Extended Data Fig. 1c). In the above analyses, we were not able to account for the migration of individuals before the diagnosis of lung cancer. As such, we analysed samples from a group of female patients with LCINS (92% LUAD, n = 228) from British

Columbia, Canada. For this dataset, $PM_{2.5}$ cumulative exposure was individually calculated for each individual through a detailed residential history from birth to current address. Most of the patients in this group (83%) were born outside Canada, and 46.7% harboured an EGFR mutation. An analysis of 3-year and 20-year $PM_{2.5}$ cumulative exposure (Methods) revealed that the frequency of EGFR-driven lung cancer cases was significantly higher after 3 years of high air pollutant exposure compared with low exposure (EGFR mutation frequency in high compared with low pollution (Methods): 73% versus 40%, respectively, P = 0.03; Extended Data Fig. 1d,e). Of note, this association was not observed after 20 years of high compared with low cumulative exposure (high versus low pollution: 50% versus 38%, respectively, P = 0.35; Extended Data Fig. 1d,e). This result could indicate that 3 years of high $PM_{2.5}$ exposure may be sufficient for EGFR-driven lung cancers to arise. To explore whether 3 years of cumulative $PM_{2.5}$ exposure is associated with lung cancer in an independent cohort not restricted to EGFR-driven cases, we obtained data from 407,509 participants in the UK Biobank. An analysis that included all participants, regardless of changes in residential location before registration, demonstrated that $PM_{2.5}$ levels (calculated at 1 μg m^3 increments) were associated with lung cancer incidence (hazard ratio (HR) = 1.08 (95% confidence interval: 1.04–1.12), raw P ≤ 0.001, false discovery rate (FDR) = 0.001; Supplementary Table 4), a result that is consistent with a previous analysis[18]. By contrast, lung cancer incidence was not associated with outdoor radon levels (HR = 0.96 (0.89–1.03), P = 0.262; Methods). Interaction tests between ever-smoking status and $PM_{2.5}$ exposure levels suggested that smoking and high $PM_{2.5}$ levels may have a combined effect on lung cancer risk (P = 0.049; Supplementary Table 4). We also noted nominal significance (raw P < 0.05, FDR > 0.05) for lip and oropharyngeal cancer (HR = 1.10 (1.01–1.19), raw P = 0.023, FDR = 0.215) and mesothelioma (HR = 1.11 (1.00–1.24), raw P = 0.048, FDR = 0.339; Supplementary Table 4 and Methods). Finally, we restricted our analysis to participants resident at the same address in the 3 years before registration (n = 371,543). This analysis showed that the relationship between lung cancer incidence and $PM_{2.5}$ exposure levels remained significant (Fig. 1d; HR = 1.07 (1.03–1.11); P ≤ 0.001). Collectively, these data, combined with published evidence[6], indicate that there is an association between the estimated incidence of EGFR-driven lung cancer and of $PM_{2.5}$ exposure levels and that 3 years of air pollution exposure may be sufficient for this association to manifest.

PM-mediated promotion of lung cancer

Next we used genetically engineered mouse models of LUAD to functionally examine whether PM exposure promotes lung tumour development. We induced expression of oncogenic human $EGFR^{L858R}$ in mouse lung through the intratracheal delivery of adenoviral-CMV-Cre to mice engineered with $Rosa26^{LSL\text{-}tTa/LSL\text{-}tdTomato}$;$TetO\text{-}EGFR^{L858R}$ alleles (ET mice). Mice were exposed to physiologically relevant doses of fine PM or PBS for 3 weeks after the induction of $EGFR^{L858R}$, and tumour burden was assessed 10 weeks after oncogene induction (Fig. 2a and Methods). In this model, rare, sporadic lung epithelial cells expressing oncogenic EGFR expanded to form pre-invasive neoplasia by 10 weeks (Fig. 2a,b). An analysis of ET mice at 10 weeks after exposure to PM revealed a dose-dependent increase in the number of pre-invasive neoplasias (PBS compared with 5 μg

PM, P = 0.047; PBS compared with 50 μg PM, P = 0.0007; Fig. 2b). PM also enhanced the number of pre-invasive neoplasia when $EGFR^{L858R}$ induction was restricted to alveolar type II (AT2) cells using lineage-specific adenoviral-SPC-Cre (Extended Data Fig. 2a). Exposure to PM before the CMV-Cre-mediated induction of $EGFR^{L858R}$ also resulted in an increased number of early neoplasias (P = 0.024; Extended Data Fig. 2b), which indicated that PM exposure before or after oncogene induction is sufficient to promote carcinogenesis.

PM exposure also increased the number of adenocarcinomas in a more aggressive CCSP-rtTa;TetO-EGFRL858R model of doxycycline-inducible LUAD (P = 0.032; Extended Data Fig. 2c). Moreover, the number of hyperplasias in an adenoviral–CMV-Cre KrasG12D model of lung cancer was also increased (Rosa26LSL-tdTomato/+;KrasLSL-G12D/+ mice; 5 μg PM, P = 0.048; 50 μg PM, P = 0.0087; Extended Data Fig. 2d). Together, these data suggest that PM can promote tumour progression in both oncogenic Kras and EGFR models of LUAD. Next we explored the mechanisms by which PM might promote EGFR-driven lung tumorigenesis. Spatial analysis of clonal dynamics throughout early tumorigenesis in ET mice indicated that the expansion of EGFR mutant cells did not occur during PM exposure but manifested in the period after PM cessation (3 weeks, P = 0.82; 10 weeks, P = 0.013; Fig. 2c,d and Methods). Both the fraction of EGFRL858R cells that grew into clusters and the number of cells within these clusters were increased in PM-exposed ET mice at 10 weeks but not at 3 weeks (Fig. 2d,e). These data suggest that PM acts in two ways to promote early tumorigenesis: by increasing the number of EGFR mutant cells with the potential to form a tumour and by increasing the proliferation rate of EGFR mutant cells within these early lesions. To test whether PM induces DNA mutagenesis, we performed whole-genome sequencing on tumours from ET mice exposed to PM or PBS. We did not observe an increase in the number of mutations in tumours from PM-exposed mice (P = 0.30; Extended Data Fig. 3a–c), or an enrichment in any established SBS signatures (P = 0.26–0.68). This result suggests that short-term exposure to PM does not enhance mutagenesis. Most of the mutations in tumours from PM-exposed mice and PBS-treated mice were attributable to the ageing SBS signatures (Extended Data Fig. 3d). We next examined whether the immune system is required for PM-enhanced EGFR-driven tumorigenesis. We crossed Rosa26LSL-tTa;TetO-EGFRL858R mice with $Rag2^{-/-}$;Il2rg−/− mice, which lack T cells, B cells and natural killer cells and have an altered myeloid compartment[19], to generate immune-deficient EGFR mutant mice after Cre delivery ($Rag2^{-/-}$;$Il2rg^{-/-}$;$Rosa26^{LSL\text{-}tTa/+}$;TetO-$EGFR^{L858R}$). In contrast to the ET mice (Fig. 2b), exposure to PM following $EGFR^{L858R}$ induction in these immune-deficient mice did not result in increased neoplasia. This result suggests that a competent immune system is required for PM-enhanced EGFR-driven lung tumorigenesis (P = 0.879; Fig. 2f). The inhalation of toxic particles induces a local response in the lung, which is mediated by macrophages and lung epithelial cells[20]. We profiled the acute myeloid response to PM in immunocompetent lungs from $Rosa26^{LSL\text{-}tdTomato/+}$ mice (control) and from mice harbouring an EGFR mutation (ET mice) 24 h after the final exposure to PM. We observed an increase in the proportion of interstitial macrophages (control mice, P = 0.043; ET mice, P = 0.034; Fig. 2g) and an increase in PD-L1 expression on these cells in both control and ET mice following PM exposure (control mice, P = 0.031; ET mice, P = 0.0061; Fig. 2g,h). No change was observed in alveolar macrophages (Extended Data Fig. 4a). In

addition, lungs from control mice displayed an increase in neutrophils, whereas dendritic cells were increased in lungs from ET mice (Extended Data Fig. 4a). Immunofluorescence staining of lungs from ET mice with the pan-macrophage marker CD68 revealed a greater density of macrophages after PM exposure, both acutely (24 h after exposure) and at 7 weeks after exposure (3 weeks, $P \leq 0.0001$; 10 weeks, $P = 0.022$; Extended Data Fig. 4b). These macrophages were confirmed to be $CD11b^{+}CD68^{+}$ interstitial macrophages (Extended Data Fig. 4c). We also observed an increased number of macrophages in both the doxycycline-inducible $EGFR^{L858R}$ mice and $Rosa26^{LSL\text{-}tdTomato/+}$; $Kras^{LSL\text{-}G12D/+}$ mice 10 weeks after induction and PM exposure (Extended Data Fig. 4d,e). These data support the hypothesis that transient PM exposure is associated with enhanced and sustained lung macrophage infiltration beyond the period of PM exposure.

PM-mediated AT2 cell reprogramming

To investigate the effects of PM exposure on early tumorigenesis, lung epithelial cells were purified by flow cytometry, and RNA sequencing (RNA-seq) was performed acutely following exposure in four different conditions: control mice exposed to PM or to PBS, and ET mice exposed to PM or to PBS. Principal component (PC) analysis of gene expression values showed that PM exposure accounted for 19% of the variance (genes differentially expressed between control mice that were exposed to PM and control mice that were exposed to PBS display higher PC2 ranks, $P < 0.001$) and EGFR mutation accounted for 38% of the variance (genes differentially expressed between ET mice that were exposed to PM and control mice that were exposed to PM display higher PC1 ranks, $P < 0.001$; Fig. 3a and Supplementary Table 5). Gene set enrichment analysis of PM-treated ET mice showed that compared with PBS-treated ET mice, the IL-6–JAKS–TAT pathway, inflammatory responses and the allograft rejection pathway were only upregulated following PM exposure in epithelium with EGFR mutations. This was in contrast to the pathways induced by PM exposure in control mice (Extended Data Fig. 5a,b). In ET mice, PM exposure led to an upregulation of genes known to regulate macrophage recruitment, including those that encode interleukin-1β (IL-1β), GM-CSF, CCL6 and NF-κB and the epithelial-derived alarmin IL-33 (Fig. 3b). AT2 cells are a probable cell of origin of lung adenocarcinoma[21], and the bleomycin lung injury model has identified a keratin 8-positive ($KRT8^{+}$) subset of these cells as progenitors that mediate alveolar regeneration driven by inflammatory signals such as IL-1β[22]. Consistent with our data showing that PM can promote tumorigenesis in $EGFR^{L858R}$ AT2 cells, we noted upregulation of genes previously associated with AT2 progenitor cell states (Fig. 3b). Deconvolution of our bulk RNA-seq expression data with signals trained on a single-cell RNA-seq dataset of bleomycin-treated mouse lungs[23] identified an increased AT2 activated progenitor score only in ET mice exposed to PM (Extended Data Fig. 5c and Methods). This result suggests that $EGFR^{L858R}$ AT2 cells are transcriptionally reprogrammed to this progenitor cell state following PM exposure. We compared the mouse RNA-seq data to a human clinical crossover study in which lung brushings from individuals who never smoked were taken after exposure to diesel exhaust and filtered air[24,25] (Extended Data Fig. 5d). A number of significantly upregulated genes within the mouse lung epithelium were also upregulated in human lung epithelium (but not reaching significance) after PM exposure, including markers of macrophage recruitment

(IL1B and IL1A) and markers of AT2 progenitor state (ORM1 and LRG1) (Extended Data Fig. 5e and Supplementary Table 5). These results identify PM-induced transcriptional changes in lung epithelium associated with inflammation and lung progenitor cell states[22]. To test whether these alterations are reflected in functional differences in epithelial cell progenitor behaviour, we isolated lung epithelial cells from ET mice following in vivo exposure to PM, and cultured them in a 3D lung-organoid-formation assay with lung fibroblasts[26] ex vivo (Fig. 3c). Non-recombined (EGFR wild-type) cells from ET mice exposed to PM did not display an increase in organoid-formation efficiency (OFE; $P = 0.075$; Fig. 3d). By contrast, recombined tdTomato^{+}EGFR^{L858R} cells demonstrated an increase in OFE ($P = 0.025$; Fig. 3d). To validate whether AT2 cells specifically are functionally altered by PM, we purified AT2 cells from non-induced ET mice and control mice exposed to PM, induced recombination in vitro[27] and then plated the cells (Extended Data Fig. 5f). Increased OFE was observed only in tdTomato^{+}EGFR^{L858R} AT2 cells from mice exposed to PM in vivo ($P = 0.0043$; Extended Data Fig. 5g,h). This result is consistent with our in vivo data (Extended Data Fig. 2a,b) and demonstrates that reversing the temporal order of oncogenic mutation initiation and PM exposure also increases OFE. PM-exposed AT2 organoids were KRT8^{+} and SPC^{+}, consistent with an AT2 progenitor state (Extended Data Fig. 5i). These data suggest that the combination of in vivo exposure to PM and induction of the EGFR^{L858R}-driver mutation increases AT2 cell progenitor function, a phenotype that is not seen with PM exposure or expression of EGFR^{L858R} alone.

PM induces IL-1β production from macrophages

We proposed that lung macrophages, which release inflammatory cytokines when exposed to PM[28], may be central to tumour promotion. We isolated AT2 cells from ET mice not exposed to PM, induced EGFR^{L858R} expression ex vivo and co-cultured the cells with macrophages exposed in vivo to either PM or PBS (Fig. 3e). Both PM-exposed interstitial macrophages and alveolar macrophages increased the OFE of EGFR mutant AT2 cells (interstitial, $P = 0.0095$; alveolar, $P = 0.0002$; Fig. 3f). This result indicates that a key mediator of PM-induced inflammation arises from macrophages. Previous reports have shown that IL-1β, derived from lung macrophages, is required for the formation of KRT8^{+} AT2 progenitor cells after bleomycin injury[22]. Therefore, we reasoned that IL-1β may be a key molecular mediator of tumour promotion and the pollutant-driven change in cell state. IL-1β was upregulated in PM-treated lungs and predominantly appeared within CD68^{+} macrophages (Extended Data Fig. 5j,k). Furthermore, treatment of EGFR mutant AT2 cells in vitro with IL-1β resulted in larger KRT8^{+}SPC^{+} organoids (Extended Data Fig. 5l). Finally, to test the requirement of IL-1β in PM-enhanced adenocarcinoma formation, we initiated oncogene expression in the doxycycline-inducible CCSP-rtTa;TetO-EGFR^{L858R} model and exposed mice to PM with concomitant administration of an anti-IL-1β or a control antibody (200 μg per dose; Extended Data Fig. 5m). Treatment with an anti-IL-1β antibody during PM exposure was sufficient to attenuate EGFR-driven LUAD formation ($P = 0.034$; Fig. 3g). Collectively, these data establish that PM-exposed macrophages are sufficient to induce a progenitor-like state in EGFR mutant AT2 cells. Moreover, macrophages are a key source of IL-1β in response to PM and IL-1β signalling is required for the promotion of PM-mediated EGFR-driven LUAD.

Oncogenic mutations in healthy lung

The model of tumour initiation and promotion is contingent on histologically normal tissue cells harbouring oncogenic driver mutations[1]. In 15 reported studies involving deep sequencing of human histologically normal tissues from different anatomical sites (n = 9,380 samples from 380 patients), an oncogenic $EGFR^{L858R}$ mutation was only reported in a single clone from a skin microbiopsy, which indicated that these mutations are rare (Supplementary Table 6). Using digital droplet PCR (ddPCR) and duplex sequencing (Duplex-seq), we sought for evidence of EGFR-driver mutations in non-cancerous lung tissue from patients with lung cancer or with cancers of other organs and from individuals with no evidence of cancer (Extended Data Figs. 7 and 8a and Supplementary Table 7). We selected healthy lung tissue from 195 out of 1,346 prospectively recruited treatment-naive patients with lung cancer from the TRACERx cohort (NCT01888601), balancing the cohort for sex, EGFR mutation status and smoking status within the limits of tissue availability (Supplementary Table 7 and Extended Data Figs. 7 and 8a,b). We used ddPCR to detect the presence of five oncogenic EGFR driver mutations (exon 19 deletion, G719S, L858R, L861Q and S768I)[29] in these tissue samples. We filtered out occurrences where the same mutation was identified in both tumour and non-cancerous tissue using MiSeq-based analysis of corresponding primary tumour tissue (Methods), which were potentially attributable to contamination from the tumour. After this filtering step, 38 out of 195 (19%) patients harboured activating EGFR mutations in healthy lung tissue that were not detectable in tumour tissue (Fig. 4a and Extended Data Fig. 8b). In one patient (identifier CRUK267), both $EGFR^{L858R}$ and $EGFR^{L861Q}$ were detected in healthy lung, but only $EGFR^{L861Q}$ (the less common driver mutation) was found in the tumour. These findings indicate that EGFR-driver mutations can be present in histologically normal lung tissue, even in patients in whom the same mutations were not selected during NSCLC tumorigenesis. To examine whether EGFR mutations exist in healthy lung tissue from people who never develop lung cancer in their lifetime, we profiled 59 healthy lung samples collected at autopsy (median 3 samples per patient, n = 19 patients) from participants in the PEACE study (NCT03004755) who died of other cancers (Supplementary Table 7 and Extended Data Figs. 7 and 8a). An EGFR-driver mutation was detected in the healthy lung of 16% (3 out of 19) patients (Fig. 4a). Despite spatially separated multiregion ddPCR analysis of healthy tissue in 15 out of the 19 patients, EGFR-driver mutations were only detected in 1 region per patient. Based on the frequency of oncogenic EGFR-driver mutations identified in healthy tissue across all patients in the PEACE and TRACERx cohorts (Supplementary Table 7), we used Bayesian inference (Methods) to estimate the presence of an EGFR-driver mutation in lung cells. The calculation showed that 1 in 554,500 lung cells (95% credible interval of 1 in 341,500 to 1 in 865,750 cells) would harbour an oncogenic EGFR mutation. We next used the TRACERx cohort to address whether there was an association of oncogenic EGFR mutations within non-cancerous tissue and exposure to ambient pollution. Anthracosis, determined by the presence of anthracotic pigment (Extended Data Fig. 8c), can act as a surrogate marker of exposure to ambient air pollution[30]. We classified anthracosis within the samples of non-cancerous lung tissue with and without EGFR-activating mutations (Fig. 4b). Although there was no association between the presence of an EGFR-driver mutation in non-cancerous tissue and anthracosis (P = 0.39; Fig. 4b), there was an association between

anthracosis and increased variant allele frequencies (VAFs) of EGFR-driver mutations (t-test P = 0.015; Fig. 4c). Although there was a trend towards enrichment of smokers in the anthracosis-positive group (Fisher's exact test, P = 0.065), several reports[30–32] have shown that cigarette smoking is not a risk factor for anthracosis. In our cohort, the degree of anthracosis observed in never-smokers and smokers did not differ, which is in line with these reports (P = 0.43; Extended Data Fig. 8d). Even though there are multiple environmental contributors to anthracosis[30], these data suggest that pollutants are not associated with the frequency of activating oncogenic mutations but rather with the expansion of EGFR mutant clones. Smoking status, sex, anthracosis and age of patients in the TRACERx cohort were entered into a multivariable model for the likelihood of anEGFR mutation in healthy tissue. Female sex demonstrated the strongest association (P = 0.06; Extended Data Fig. 8e). We next addressed whether driver mutations existed at other genomic loci in EGFR and in KRAS using an independent ultradeep sequencing platform in a separate group of patients with and without cancer (n = 81). We analysed 48 samples of non-cancerous lung tissue from the PEACE study (lung cancer, n = 9; other cancer, n = 39) and 33 samples of healthy lung tissue derived from the Biomarkers and Dysplastic Respiratory Epithelium (BDRE) study (NCT00900419; Supplementary Table 7 and Extended Data Figs. 7 and 8a). The BDRE cohort consisted of patients with suspicious lung nodules identified through computed tomography scans and who were referred for evaluation by navigational bronchoscopy. For each patient, a brushing sample enriched for bronchial epithelial cells (>89%)[33,34] from the uninvolved contralateral lung was taken for research purposes and used as the source of healthy tissue. Profiling was carried out using Duplex-seq, which covers a broader range of mutations (EGFR exon 18, 19, 20 and 21, KRAS exon 2 and 3 and other cancer genes). Thus, we only considered mutations featured in the cancer gene census[35] and further filtered these by evidence of driver mutation status in the literature (Supplementary Table 8). In 24 out of 68 cancer cases for which tissue was available, we also performed Duplex-seq or MiSeq on the corresponding tumour tissue to confirm that the identified mutations were found exclusively in the healthy tissue samples. Based on the Duplex-seq data, 13 out of 81 (16%) samples harboured an EGFR-driver mutation (E709X, G719X, T725M, exon 19 deletion, R765X, R776X, L858R or L861X; Fig. 4d and Extended Data Fig. 9a), whereas 43 out of 81 (53%) samples harboured a KRAS driver mutation (G12X, G13X or Q61X; Fig. 4d and Extended Data Fig. 9b). BRAF inhibitors used to treat BRAF mutant melanomas are known to promote the accelerated growth of clones harbouring RAS mutations[36]. Excluding patients with melanoma from the analysis did not change the percentage of cases harbouring a KRAS mutation (36 out of 68 (53%); Extended Data Fig. 9c), which suggests that this parameter did not confound our analysis. Concordant with KRAS being commonly mutated in ever-smoker LUAD, KRAS mutation frequency and VAFs were significantly higher than EGFR mutation and VAFs in samples from ever-smokers (P = 0.012; Extended Data Fig. 9d). Moreover, VAFs of high-confidence KRAS mutations were consistently higher than those in EGFR in the four ever-smoker cases that harboured oncogenic mutations in both genes (P = 0.015; Extended Data Fig. 9d), which indicates that KRAS mutant clones may be more highly selected for than EGFR mutant clones in healthy lungs of ever-smokers. In summary, 54 out of 295 (18%) of samples of non-cancerous lung tissue harboured an EGFR driver mutation, and 43 out of 81 (53%) samples of non-cancerous lung tissue harboured a KRAS driver mutation. No associations

between EGFR or KRAS mutation in non-cancerous tissue and smoking status or cancer diagnosis were observed (Supplementary Table 9). To address whether oncogenic mutations accumulate with the natural ageing process, we examined the driver mutation frequency in all 31 genes (including EGFR and KRAS) present in the Duplex-seq panel. We limited this analysis to 17 never-smoker individuals from the PEACE study to control for any effect of smoking. Consistent with previous work[37,38], there was a significant correlation between age and mutation count (Fig. 4e).

Discussion

In this study, we explored the paradigm of tumour promotion driven by the air pollutant PM in the development of lung cancer. We build on previous studies proposing that engine exhaust[39] and air pollution[40] induce lung tumours through genotoxicity, induction of oxidative stress and inflammation. We propose that PM can trigger the expansion of pre-existing mutant lung cells through an inflammatory axis that may be amenable to therapy to limit the risk of tumour promotion. Extending previous findings that established associations between air pollution and lung cancer[18,41], including in LCINS[6], we found an association between the frequency of EGFR mutant lung cancer incidence and increasing $PM_{2.5}$ levels. Temporal analysis suggested that 3 years of $PM_{2.5}$ exposure may be sufficient to increase the risk of developing EGFR-driven lung cancer. A limitation of our epidemiological analysis is its ecological nature: using aggregate data instead of participant-level data. We also acknowledge that variables such as female sex, Asian ancestry and adenocarcinoma histology, which are associated with EGFR mutation status, may confound our conclusions. We balanced our study cohorts with respect to sex and covered geographically and ethnically distinct populations, and when restricting the analysis to LUAD in the English cohort, the positive association remained significant. This study suggests that PM exposure contributes to the observed geographical disparities of EGFR-driven lung cancer, in addition to other established intrinsic (for example, germline genetics[5]) and extrinsic (for example, occupational exposure[3]) factors, and it will be important to understand how these factors interact to increase risk. We observed that PM induces an altered progenitor state in EGFR mutant AT2 cells through the macrophage release of IL-1β, which promotes lung cancer. A caveat of our work is that these mouse models will develop cancers in the absence of PM and probably do not replicate the complex spectrum of mutations found in healthy tissue of a healthy adult. However, they provide controlled environments to provide insight into early tumorigenesis. These experiments demonstrate that a key driver of tumorigenesis is a clinically targetable inflammatory axis that could be applicable to a range of risk factors and malignancies[15,42]. It is notable that the antibody canakinumab, which is targeted against IL-1β, a cytokine induced in both mice and humans following PM exposure, has been shown to reduce lung cancer incidence in the cardiovascular prevention trial CANTOS[43].

A limitation of our DNA profiling strategies of non-cancerous tissue is that we did not purify epithelial cells, specifically AT2 cells, which are the probable initiators of lung tumours. Further work would be required to pinpoint which lineages harbour these mutations. From histological analyses, AT2 and AT1 cells account for on average 22% of distal lung parenchyma cells in autopsy or surgical resection lung samples, mixed with 37% endothelial cells, 37% interstitial cells and 3% macrophages[44]. Our results provide additional evidence

that a major trigger of cancer development is not only the inevitable acquisition of driver mutations in healthy epithelium but also intrinsic and extrinsic mechanisms that promote nascent mutant cell expansion and progenitor activity. Assuming little can be done to prevent the acquisition of oncogenic mutations with age, it may be beneficial to address whether additional carcinogens promote cancer through similar inflammatory mechanisms. Broad approaches will be necessary to establish how these carcinogens, as well as potential hormonal, environmental and germline influences, might promote or restrict mutant clone expansions and contribute to tumour promotion. There is an urgent need for carcinogenic assays to identify potential tumour-promoting agents across different tissues and to understand tissue-specific mediators. Such efforts may guide new screening paradigms in high-risk, under-served populations and molecularly targeted cancer prevention approaches to inhibit cancer initiation. In conclusion, our data suggest a mechanistic and causative link between air pollutants and lung cancer, as previously proposed[45], and substantiate earlier findings on tumour promotion[1], providing a public health mandate to restrict particulate emissions in urban areas.

Methods

ddPCR of tumour and healthy lung tissue samples from the TRACERx and PEACE studies

This project leverages the infrastructure established by the national pan-cancer research autopsy programme (PEACE, NCT03004755) and the prospective, longitudinal cohort study (TRACERx) of NSCLC (NCT01888601)[12].

To explore whether clinical disparities in lung cancer in never-smokers were reflected in EGFR mutation status in non-cancerous lung tissue, we sought to assemble a cohort comprising participants in the TRACERx study that was as best as possible balanced for sex (male individuals compared to female individuals), smoking status (never-smoker compared with ever smoker) and EGFR mutation status in tumour samples (EGFR mutation versus EGFR wild-type). To uncover whether EGFR mutations were also found in non-cancerous lung tissue from patients who never acquire a lung cancer diagnosis in their lifetimes, we also assembled a cohort of individuals from the PEACE study.

Based on tissue that was available for study, our dataset consisted of 195 tumour and 195 non-cancerous lung tissues from 195 patients from the TRACERx study and 59 non-cancerous lung tissues from 19 participants in the PEACE study (median 3 samples per patient, range of 1–10).

For the TRACERx study, tumour and non-cancerous lung tissue were obtained at surgery. Healthy (non-cancerous) lung tissue was collected distally from the primary tumour tissue (at least approximately 2 cm apart). All tissue was initially snap-frozen and then a portion fixed and made into a formalin-fixed paraffin-embedded (FFPE) block. A haematoxylin and eosin (H&E) section of each block underwent pathology review. DNA was extracted from frozen healthy and tumour tissue proximal to these sections. For the PEACE study, healthy lung tissue was collected at post-mortem tissue from patients who never acquired lung cancer in their lifetimes. Each piece of collected tissue was immediately bisected, and one half was snap-frozen and the other was fixed and made into a FFPE block. The H&E section

of each block underwent pathology review. DNA was then extracted from an adjacent frozen healthy tissue sample.

All aforementioned H&E slides from tissues underwent central pathology review. In particular, to exclude the possibility of contamination with tumour cells, thoracic pathologists confirmed that all healthy lung tissue samples did not contain any indication of tumour tissue or morphologically defined, pre-invasive disease. Thoracic pathologists also identified anthracotic pigment and reflected this in a binary score for its presence. For anthracosis-positive cases, the proportion of the tissue covered by anthracotic pigment was quantified.

EGFR mutation profiling of non-cancerous tissue samples by ddPCR

DNA was extracted from healthy lung tissue samples as previously described[12]. The DNA concentration was measured using Qubit, and up to 3,000 ng of DNA was fragmented to approximately 1,500 bp using a Covaris E220 evolution focused-ultrasonicator following the manufacturer's standard protocol. SAGAsafe assays[46] for five EGFR target variant alleles (L858R, exon 19 deletion, S768I, L861Q and G719S) were used (SAGA Diagnostics). SAGAsafe is a digital PCR-based ultrasensitive mutation detection technology that utilizes an alternative chemistry alongside a modified thermocycling program, such that the true positive variant allele signal is enriched during a linear phase, and signals for both the variant and the wild-type alleles are amplified during the exponential phase. The method effectively suppresses the false-positive variant allele signal arising from polymerase base misincorporation errors and DNA damage, making reliable detection of rare-event mutations possible to exceedingly low limits of detection. The assays were performed on a Bio-Rad QX200 Droplet Digital PCR system. At least three positive droplets were required to call a sample positive. Using control experiments containing 265,000–381,000 copies of wild-type genome equivalents per test, the achievable limit of detection for the five EGFR SAGAsafe assays was determined to be at least 0.004% VAF. For each patient sample, 500 ng of fragmented DNA (corresponding to about 150,000 copies of genome equivalents) was analysed per assay across 4 reaction wells, with positive and negative control samples included for every run.

The copy number concentration of the variant and the wild-type alleles was calculated as follows:

$$C_{V_i} = \frac{-\ln(1 - \frac{P}{T})}{V_d} \times \frac{V_r}{V_i}$$

where C_{Vi} is the copy number concentration of the target (variant or wild-type allele) in the input DNA sample, P is the number of positive droplets for the target, T is the number of total droplets analysed, V_d is the volume a droplet (0.85 x10^{-3} μl), V_r is the total volume of a ddPCR reaction (20 μl), and V_i is the input volume per ddPCR reaction of the input DNA sample.

The VAF was calculated as follows:

$$VAF = \frac{C_{V_i}^{\text{Variant}}}{C_{V_i}^{\text{Variant}} + C_{V_i}^{Wild-type}} \times 100\,\%$$

To estimate the EGFR mutation rate, we considered all five oncogenic EGFR mutations detected by ddPCR in all TRACERx and PEACE samples analysed (253 samples in total). Using the approximate Bayesian computation model, we simulated ddPCR results of oncogenic EGFR mutations and inferred a mutation rate of 4.07 $x10^{-7}$ per mutation (confidence interval: 1.61 x 10^{-7} to 6.08 x 10^{-7}). Considering this mutation rate, we estimated that the frequency of identifying 1 EGFR mutation (of any of the 5 mutation types) would be 1 in 2,035,000 (95% confidence interval: 1 in 805,000 to 1 in 3,040,000). When we took the average of the two limits of the confidence interval, we obtained an estimate of an EGFR mutation being present in 1 in 554,500 cells (or around 1:600,000 cells).

EGFR mutation profiling in corresponding tumour tissue by MiSeq

To exclude the presence of clonal or subclonal spatially distinct EGFR mutations that may be present in the corresponding matched lung tumour, we performed multiregion deep next-generation sequencing of NSCLC samples from the same patients (>3,000x coverage) of 19 driver genes (including EGFR) using the MiSeq platform. We sequenced 751 tumour regions from the 195 tumours (median of 3 regions per tumour) with an achievable limit of detection in each tumour region of 0.966% based on a median sequencing depth per region of 3,490x and a MiSeq error rate of 0.473%[47].

For each tumour region and matched germline, capture of a custom panel of genes (including the EGFR locus) was performed on 125 ng DNA isolated from genomic libraries. The TruSeq Custom Amplicon Library Preparation method was used. Following cluster generation, samples were 100 bp paired-end multiplex sequenced on an Illumina MiSeq platform at the GCLP Laboratory at University College London, as previously described[12]. The generated data were aligned to the reference human genome (hg19). Mutations were called as previously described[12].

Duplex-seq of samples from the PEACE and BDRE studies

Non-cancerous lung tissue samples—Samples from the PEACE cohort were collected as described above. For Duplex-seq, we obtained additional non-cancerous lung tissue from 48 participants of the PEACE study. Here patients with lung cancer or with another cancer type were profiled (lung cancer, n = 9; other cancer, n = 39).

Participants in the BDRE study (NCT00900419) consisted of individuals recommended for a computed tomography (CT) scan based on age, smoking history or other indications. If a suspicious nodule was detected by CT scan, a navigational bronchoscopy was indicated. The nodule site was sampled for accurate diagnosis. For each patient, a brushing from a remote site in a contralateral lobe was also taken for research as a representative sample of non-cancerous tissue and subsequently profiled for mutations using Duplex-seq. The absence of nodules or masses detected by chest CT scans was indicative of the non-tumour nature of

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

these contralateral samples. Each procedure was performed under fluoroscopic guidance, with the brush advanced from the sheath only after documentation that the working channel was in the peripheral airways.

EGFR and KRAS mutation profiling by Duplex-seq—Genomic DNA was extracted from brushing samples using a Qiagen DNeasy Blood and Tissue kit according to the manufacturer's instructions. Duplex libraries were prepared using a commercially available kit from Twin-Strand Biosciences (CKD-00042 panel 000323), starting with 250 ng of input DNA. Custom probes were designed for targeted capture of EGFR exons 18, 19, 20 and 21, and KRAS exons 2 and 3, along with 29 other cancer genes.

By independently capturing and sequencing the two strands of DNA for selected genomic regions, combined with the use of a common unique molecular identifier for both strands, Duplex-seq enables the detection of rare mutations[46] with a sensitivity of less than 1 in 10^7. After shearing and capturing of gDNA spanning the panel, primers were ligated so that the two strands of DNA for each segment were uniquely labelled and matched with its opposing strand. These strands were then amplified, and libraries were sequenced on a NovaSeq 6000 sequencing system (Illumina), and sequencing data were processed using a DNAnexus platform. Samples had an average number of 150,000,000 raw reads, producing a mean on-target duplex depth of 4,500. Duplex-seq reads were processed using a previously published pipeline[48], similar to a bioinformatics pipeline provided by TwinStrand BioSciences. Using this, we were able to identify mutations that were present in both the involved and contralateral lung samples.

Epidemiological studies

UK Biobank dataset

The UK Biobank (UKBB) study comprises more than 500,000 participants, aged between 37 and 73 years, who were recruited between 2006 and 2010. Participants provide detailed information regarding a comprehensive set of lifestyle factors, in addition to physical measurements and biological samples. PM air pollution levels (in 2010) were estimated for addresses within 400 km of the Greater London monitoring area using a land-use regression model developed as part of the ESCAPE study[49].

Lung cancer cases were those with International Classification of Diseases (ICD; tenth revision) codes C33 or C34. Associations between $PM_{2.5}$ levels and lung cancer incidence in the UKBB data have already been calculated and previously reported[18].

We accessed the UKBB data under project number 82693. Ethical approval of the UKBB study was 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.

To impute missing data, we first excluded all participants who had any cancer diagnosis pre-recruitment, or a cancer diagnosis date entry but no corresponding cancer annotation, alongside those with missing particulate matter or genetic principal components data.

Multiple imputation with chained equations[50] was used to impute missing smoking status (categorized 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), body–mass index (BMI) (<1% missing), household income (dichotomized by ≥GBP£31,000 annually; 14.6% missing) and educational attainment (split by degree or professional qualification status; 1.31% missing) values. In addition to these variables, imputation models used the following variables to predict values for missing data: $PM_{2.5}$, age at baseline, sex, BMI and the first 15 genetic PCs (to account for ethnicity). These were used alongside cancer outcome and duration of follow-up. We used predictive mean matching, logistic regression and random forest for continuous, binary and categorical variables, respectively, performing a maximum of 180 iterations for the generation of each imputed dataset. This produced 15 complete versions of the original dataset in which the missing values were imputed. This dataset comprised 407,509 individuals and represented 28 cancer types. Each imputed dataset was independently used in the same analysis protocol.

Participants were followed up from recruitment until either date of each cancer diagnosis (obtained through linkage to national cancer registries) or censoring, which was defined as time of death, lost to follow-up or the end of 2018, whichever was earlier. We created a multivariate Cox regression model for each imputed dataset and primary cancer type with ≥100 cases (excluding non-melanoma skin cancer, and cancers restricted to a single sex), and pooled results across these models, which were consistent for each cancer type, into a single set using Rubin's rules[50]. Confidence intervals were calculated using $e^{estimate_{pooled}+-(1.96xstandarderrorpooled)}$. These models included the same covariates as in the imputation model. For laryngeal alongside lip and oropharyngeal cancers, we further corrected for alcohol consumption, excluding those participants with missing alcohol data owing to the high missingness of these variables (30.7%). Schoenfeld residuals were examined to assess the proportional hazards assumption, with non-proportionality confirmed using Kaplan–Meier curves for binary and categorical variables. Potential departures from the proportional hazards assumption were noted for anal (smoking status), bladder (genetic PC 12), kidney (age and smoking status) and melanoma (genetic PC 9 and sex). We note high median (across all 15 imputations) variance inflation factor values (≥5) for the following covariates: genetic PC 1 (other and unspecified biliary tract parts); PC 2 (acute myeloid leukaemia, follicular nodular non-Hodgkin lymphoma, larynx, mesothelioma, other and unspecified biliary tract parts, peripheral and cutaneous T lymphomas, retroperitoneum and peritoneum); and PC 3 (acute myeloid leukaemia, follicular nodular non-Hodgkin lymphoma, larynx, mesothelioma, other and unspecified biliary tract parts, peripheral and cutaneous T lymphomas). Finally, we report FDR-corrected P values for the association between $PM_{2.5}$ levels and cancer incidence to account for multiple testing.

Our methods differed from those of Huang et al.[18] in the following ways: (1) we increased the number of imputations from 5 to 15 and iterations from 90 to 180; (2) we augmented our multivariate analysis to better account for the effect of smoking by categorizing participants into never, previous and current smokers, and included passive smoking; (3) we included the first 15 genetic PCs in our multivariable analysis of $PM_{2.5}$ and cancer incidence. An interaction test between $PM_{2.5}$ and smoking was then performed for lung cancer, considering only participants with complete covariate data in the multivariable Cox

regression. For the LUAD-specific analysis, we considered only participants with cancer registry histology entries that map to LUAD (Supplementary Table 4). Imputations and all downstream modelling was performed independently for this analysis. To take into account migration, as the $PM_{2.5}$ data are available for each participant's address, we assumed that participant $PM_{2.5}$ exposure levels remained constant throughout the study period. To account for exposure misclassification, we additionally performed a separate analysis that included only participants who had lived at their current address for at least 3 years before baseline. All imputations and downstream analysis was performed independently for this subgroup.

Radon exposure data from the British Geological Survey (BGS) was merged with the UKBB dataset based on home location coordinates. As the data from BGS had greater spatial resolution, values were aggregated by the mode radon potential class (breaking ties through taking the higher class value) across all BGS coordinate values that map to each rounded coordinate in the UKBB. Imputations and downstream analyses were performed as described above, using modal radon exposure instead of $PM_{2.5}$.

Comparison of the UKBB population with the general UK population

Estimated HRs from UKBB analyses are higher than in some population-based epidemiological surveys[41], which may reflect over-representation of less wealthy, never-smoker individuals in the UKBB. We have provided a table (Supplementary Table 4) comparing some characteristics between the UKBB population we studied and UK population estimates for reference. Compared with the general population, UKBB participants consisted of fewer current smokers, were more highly educated, had lower household income, were more likely to be female individuals, older, white and to live in areas with lower $PM^{2.5}$ levels.

Within-country datasets England dataset (NDRS)

Air pollution, lung cancer incidence and EGFR mutation status could be estimated for 20 Cancer Alliance regions in England. This was the geographical level at which all three factors could be quantified. Annual $PM_{2.5}$ air pollution data (μg m^{-3}) from 2006 to 2017 was obtained at the grid code level (1x1 km) from DEFRA[51]. Radon potential (defined as the estimated percentage of homes in an area above the radon action level) in 2011 was obtained from the British Geological Survey at the grid code level[52]. Postal code coordinates were sourced from the Office of National Statistics 2018 Postal Code Directory[53]. To link every postal code to a grid code with pollution data, the coordinates of every postal code centroid was mapped to those of the nearest grid code centroid using the RANN package in R. The postal codes with pollution data were binned into 1 of 20 Cancer Alliance regions. Then PM_2.5 concentration estimates were aggregated to the Cancer Alliance region level and then averaged over the period 2008–2017 for 2018 diagnoses, 2007–2016 for 2017 diagnoses and 2006–2018 for 2016 diagnoses—these were selected because they represented the 10 years before a lung cancer diagnosis. The air pollution levels in each Cancer Alliance region were broadly stable (within 5 μg m^{-3}) in this time period. Incidence data on 118,019 (2016, 39,229; 2017, 39,500; 2018, 39,290) lung cancers (ICD codes C33 to C34) diagnosed in England between 1 January 2016 and 31 December 2018 were extracted from the National Cancer Registration Dataset (AV2018 in CASREF01 (end of year snapshot)),

held by the National Disease Registration and Analysis Service at England's NDRS. Lung cancer incidence for each Cancer Alliance region was calculated based on these cases. This represented a predominantly white cohort: white, 92.06%; Asian, 1.48%; Chinese, 0.23%; Black, 1.05%; mixed: 0.28%; other: 0.94%; unknown: 3.96%. The age-standardized lung cancer incidence (using population counts obtained from the Office of National Statistics 2019 (2018 mid-year estimates)) was obtained according to each 5-year age group and sex. Incidences were then combined across age and sex to produce a single value for each Cancer Alliance region as follows: lung cancer incidence = (sum(wi × xi/di)/sum(wi)) × 100,000. Where wi is the European population standard, di is the population count and xi the case count. Standardized rates were standardized according to the 2013 European Standard Population. Confidence intervals for age-standardized rate point estimates were calculated using the Dobson method. For lung cancer diagnoses listed above, EGFR mutation statuses were extracted from the National Cancer Registration Dataset (AT_GENE_ ENGLAND table in the CAS2210 monthly snapshot), which includes data on somatic tests undertaken from 1 January 2016 to 31 December 2019. Only cases with ‘Overall: TS’ as ‘a:abnormal’ and ‘b:normal’ for EGFR were used in the calculation for the EGFR mutation rate (n = 25,567). The EGFR mutation rate was calculated for each Cancer Alliance region as follows: EGFR mutation rate = [number of a:abnormal]/[(number of a:abnormal) + (number of b:normal)]. The NDRS data included in this study were collected and analysed under the National Disease Registries Directions 2021, made in accordance with sections 254(1) and 254(6) of the 2012 Health and Social Care Act. Further ethical approval for this study was not required per the definition of research according to the UK Policy Framework for Health and Social Care Research.

South Korea dataset (Samsung Medical Center)

Air pollution, lung cancer incidence and EGFR mutation status could be estimated for 16 geographical regions in South Korea. This was the geographical level at which all three factors could be quantified. $PM_{2.5}$ air pollution data were obtained from Air Korea[54] for the years 2015–2017 for 16 standard geographical regions across Korea. Within each of the geographical regions, we averaged $PM_{2.5}$ levels across the 2-year period before the year of lung cancer diagnosis. $PM_{2.5}$ levels between 2015 and 2017 were broadly stable. We were only able to include $PM_{2.5}$ data for a 2-year period for 2017 and 2018 diagnoses, as air pollution data per region in Korea was only available starting from 2015. Lung cancer incidence data were obtained from the Korean National Cancer Center[55] for the years 2017 to 2018 for 16 geographical regions across Korea. Sex and smoking data were not available. Lung cancer incidence was obtained separately for each year and considered independently in Pearson correlations that are described below. Lung cancer EGFR mutation status was obtained from Samsung Medical Center lung cancer diagnoses for the years 2017 to 2018 for 16 geographical regions across Korea (n = 2,563). The EGFR mutation rate was calculated as described above. The study was conducted under an institutional review boardapproved protocol (number 2021-06-043) at the Samsung Medical Center.

Taiwan dataset (Chang Gung Medical Foundation)

Air pollution, lung cancer incidence and EGFR mutation status could be estimated for 12 standard geographical regions in Taiwan. This was the geographical level at which all

three factors could be quantified. Annual $PM_{2.5}$ air pollution data were obtained for 12 standard geographical regions in Taiwan from the Environmental Protection Administration Executive Yuan R.O.C. (Taiwan)[56]. $PM_{2.5}$ (μg m^{-3}) concentration estimates were available for each county in Taiwan from 2006 to 2017. We averaged $PM_{2.5}$ levels across the period (up to 10 years before a 2-year washout period) before the year of lung cancer diagnosis. For example, for a diagnosis in 2017, 2006–2015 aggregated air pollution levels were used for analysis, whereas for a diagnosis in 2011, 2006–2009 aggregated air pollution levels were used for analysis. A 2-year washout period was necessary to account for substantial decreases in air pollution levels after 2013. Institutional lung cancer incidence and EGFR mutation rates for each of 12 different counties in Taiwan were obtained from the Chang Gung Research Database for the years 2011–2017 (n = 4,599). Lung cancer incidence was obtained separately for each year and considered independently in Pearson correlations that are described below.

Institutional lung cancer incidence was estimated based on recorded lung cancer diagnoses in all of Chang Gung Medical Foundation hospitals, and the age-standardized rates per 100,000 were calculated using the world (World Health Organization 2000) standard population of lung cancer incidence. EGFR mutation testing data were available for all of these cases. However, only nine counties had at least ten cases with EGFR mutation tested per year and constituted >5% of the total population; these were the counties that were retained for analysis. The EGFR mutation rate was calculated as outlined above. The data from the Taiwan cohort was from the Chang Gung Research Database, which is approved by the institutional review board of Chang Gung Medical Foundation (202101202B0).

Relationship between EGFR mutant lung cancer incidence and $PM_{2.5}$

Analyses were performed separately for each of the three cohorts: England, South Korea and Taiwan. For each geographical region (for example, each country or the 20 Cancer Alliance regions in England), EGFR-driven lung cancer incidence was calculated by multiplying the total lung cancer incidence by the EGFR mutation rate (as reported as a proportion out of 1) as follows: EGFR mutation lung cancer incidence = lung cancer incidence x EGFR mutation rate. EGFR mutant lung cancer incidence values were compared with mean $PM_{2.5}$ values across geographical regions using Pearson correlation tests, weighted Pearson correlation tests (to account for number of tested cases in each geographical region) and robust linear regression (to account for outliers).

Sensitivity analysis for the England and Korea datasets

In the England dataset, there were two Cancer Alliance regions (South East London and Thames Valley) with sparse data owing to data unavailability (<5% of lung tumours diagnosed in 2016–2018 have a definitive test result recorded for EGFR). To exclude the possibility of this confounding our analysis, we performed a sensitivity analysis, whereby we excluded data from these two regions. Of note, the correlation between $PM_{2.5}$ and EGFR-driven lung cancer incidence was still significant ($r = 0.55$, $P = 0.019$) after these exclusions. Similarly, in the South Korea dataset, Jeju-do (2017) was excluded owing to poor data availability. The correlation between $PM_{2.5}$ and EGFR-driven lung cancer incidence was still significant ($r = 0.38$; $P = 0.033$) after this exclusion. However, for the

sake of completion, we report the full datasets (including these two regions in England regions and one region in South Korea region) in the main text.

Canada dataset (BC Cancer Research Centre, Vancouver BC, Canada)

This dataset comprises 228 lung cancer cases from female patients and has been previously reported[6]. These patients were seen at the Thoracic Surgery Department of the Vancouver General Hospital or the BC Cancer Vancouver Cancer Center between 15 November 2017 and 31 May 2019, and were prospectively invited to take part in the study. Detailed residential histories from birth to cancer diagnosis for residences within Canada and previous residences outside of Canada (for foreign-born immigrants) were recorded. Street and city address or postal codes enabled accurate linking of residential locations to satellite-derived $PM_{2.5}$ exposure data that were available from 1996 onwards. A personal $PM_{2.5}$ cumulative exposure value was individually calculated using a detailed residential history from birth to current address, and input into Geographical Information System mapping. By applying high-resolution (10 x 10 km) concentration estimates of $PM_{2.5}$ from satellite observations, chemical transport models and ground measurements to each individual's residential history, a cumulative exposure value was estimated by taking into account the intensity and duration of exposure and summing over all residences. EGFR mutation status for each patient was obtained from each patients' hospital record. This study was approved by the UBC_BC Cancer Research Ethics Board.

Defining pollution exposure groups

Low, intermediate and high air pollution groups were defined by considering quintiles of the distribution of $PM_{2.5}$ exposure levels across the entire dataset (3 years of cumulative pollution data and 20 years of cumulative pollution data). The following thresholds were applied: bottom quintile, 6.77 μg m^{-3}; top quintile, 7.27 μg m^{-3}; $PM_{2.5}$ low, $PM_{2.5}$< bottom quintile; $PM_{2.5}$ intermediate, $PM_{2.5}$> bottom quintile and $PM_{2.5}$<top quintile; $PM_{2.5}$ high, $PM_{2.5}$> top quintile.

Comparing EGFR mutant frequencies

EGFR mutation frequencies were compared between high and low pollution exposure groups using chi-squared tests. Two comparisons were performed: high versus low pollution (based on 3 year data) and high versus low pollution (based on 20 year data).

Preclinical studies

Animal procedures

Animals were housed in ventilated cages with unlimited access to food and water. 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. Both male and female mice aged 6–15 weeks were used. $EGFR^{L858R}$ (Tg(tet-O-EGFR*L858R)56Hev) mice were obtained from the National Cancer Institute Mouse Repository. Rosa26tTA and Rosa26-LSL-tdTomato mice were obtained from the Jackson

laboratory. Mice were backcrossed onto a C57Bl6/J background and further crossed to generate Rosa26$^{LSL\text{-}tTa/LSL\text{-}tdTomato}$;TetO-EGFR^{L858R} mice. CCSP-rtTa;TetO-EGFR^{L858R} and Rosa26$^{LSL\text{-}tTa/LSL\text{-}tdTomato}$;Kras^{LSL_G12D} mice have been previously described[57,58]. After weaning, the mice were genotyped (Transnetyx) and placed in groups of 1–5 mice in individually ventilated cages, with a 12-h daylight cycle. Cre-mediated recombination was initiated by adenoviral CMV-Cre (Viral Vector Core) delivered by intratracheal intubation (2.5 x 10^7 virus particles per 50 pl), by Ad5-SPC-Cre (Viral Vector Core, donated by A. Berns) delivered by intratracheal instillation (2.5 x 10^8 virus particles per 50 µl)[21] or by using chow containing doxycycline obtained from Harlan-Tekland. For antibody treatment, mice were given 200 µg of anti-mouse/rat IL-1β (B122, InVivoMAb, BE0246) or rat IgG control (InVivoMAb, BE0091) by intraperitoneal injection on the same day as PM exposure.

For exposure to fine PM or PBS, SRM2786 from the National Institute of Standards and Technology (obtained from Sigma Aldrich) was resuspended in sterile PBS using sonication, and the particle size distribution was confirmed using a dynamic light scattering analyser (Zetasizer, mean particle diameter 2.8 µm). SRM2786 has certified mass fraction values of both organic and inorganic constituents from multiple analytical techniques and represents fine PM from a modern urban environment[59]. Mice were briefly anaesthetized using 5% isoflurane followed by intratracheal administration of 50 µg or 5 µg in a volume of 50 µl (ref. 60). Mice were intratracheally administered with PM or PBS three times per week for 3 weeks with at least 48 h between each administration. FACS analysis and cell sorting. For flow cytometry analysis of immune cells, mouse lungs were minced into small pieces, incubated with collagenase (1 mg ml^{-1}; ThermoFisher) and DNase I (50 U ml^{-1}; Life Technologies) for 45 min at 37 ÅãC and filtered through 100 .m strainers (Falcon). Red blood cells were lysed for 5 min using ACK buffer (Life Technologies). Cells were stained with fixable viability dye eFluor780 (BD Horizon) for 30 min and blocked with CD16/32 antibody (BioLegend) for 10 min. Cells were then stained with antibodies for 30 min (Supplementary Table 10). Intracellular staining was performed using a Fixation/ Permeabilization kit (eBioscience) according to the manufacturer's instructions. Samples were resuspended in FACS buffer (2% FCS in PBS) and analysed using a BD Symphony flow cytometer. Data were analysed using FlowJo (Tree Star).

For flow cytometry sorting of AT2 cells[61], epithelial cells and immune cells, minced lung tissue was digested with Liberase TM and TH (Roche Diagnostics) and DNase I (Merck Sigma-Aldrich) in HBSS for 30 min at 37 °C in a shaker at 180 r.p.m. Samples were passed through a 100 µm filter, centrifuged (300g, 5 min, 4 °C) and red blood cells were lysed as described above. Extracellular antibody staining was then performed followed by incubation in DAPI (Sigma Aldrich) to label dead cells. Gating strategies for sorting and analysis are outlined in Extended Data Fig. 6. Cell sorting was performed on Influx, Aria Fusion or Aria III instruments (BD).

Immunohistochemistry

Mouse lungs were fixed overnight in 10% formalin and embedded in paraffin blocks. Then 4 µm tissue sections were cut, deparaffinized and rehydrated using standard methods. Antigen retrieval was performed using pH 6.0 citrate buffer and incubated with antibodies

(Supplementary Table 10). 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 IL1B RNA transcripts by RNAscope (ACD Biotechne) following the manufacturer's instructions. Staining for CD68 protein was subsequently performed and detected using OPAL fluorescence following the manufacturer's protocols (Akoya). Probes visualized through fluorescence were used to detect IL-1β RNA and CD68 protein simultaneously. Slides were imaged using a Leica Zeiss AxioScan.Z1 slide scanner. Tumour grading and lesion analysis was carried out by two board certified veterinary pathologists. EGFR mutant cell foci were quantified from cell coordinate data by clustering cell positions by density using the DBSCAN algorithm, implemented in Python with the scikit-learn library[62]. We chose an EPS value of 35 for DBSCAN clustering as this produced spatial clusters with excellent concordance to visual inspection of foci in the original histological images. To assess the fraction of clusters that had expanded, we reasoned that wild-type cells may divide only once between 3 and 10 weeks, which is based on the low proliferation rate of alveolar epithelial cells[63]. As there was an average cluster size of 2 EGFR mutant cells at 3 weeks, we defined clusters of >5 cells at 10 weeks as ‘expanded clusters’ that expanded above expected. Segmentation and analysis of immunohistochemistry and immunofluorescence images was carried out using QuPath[64].

Whole-genome sequencing

ET lung tumours from PBS-treated mice (n = 5) and PM-exposed mice (n = 5) were collected at ethical end points. Individual lung tumours were dissected from lung lobes and snap frozen. Germline DNA was extracted from t ail tissue. DNA was isolated and prepared for whole-genome sequencing (WGS), which was followed by sequencing on a NovaSeq instrument (Illumina) to achieve target coverage of 100x for PBS-treated and PM-exposed samples and 30x for germline samples. Sequences from all 20 samples were processed using the Nextflow (v.21.10.3) Sarek pipeline (nf-core/sarek v.3.0). In brief, sequences were aligned with BWA (v.0.7.17) to mm10, and mutations were called using Mutect2 (gatk4: 4.1.8.1). Only mutations labelled as ‘PASS’ by Mutect2 that were uniquely present in each tumour were considered for analysis. Mutational signatures were called using the DeconstructSigs R package[65], restricting our analysis to the following common SBS signatures: SBS1, SBS4, SBS5, SBS2, SBS13, SBS40, SBS92, SBS17a, SBS17b and SBS18.

RNA-seq

$CD45^{-}CD31^{-}TER119^{-}EpCAM^{+}$ lung cells from PBS-treated and PM-exposed mice were sorted by flow cytometry. Total RNA was isolated using a miRNeasy Micro kit (Qiagen) according to the manufacturer's instructions. Library generation was performed using KAPA RNA HyperPrep with RiboErase (Roche), followed by sequencing on a HiSeq (Ilumina) instrument to achieve an average of 25 million reads per sample. The RNA-seq pipeline of nf-core framework (v.3.3) was launched with Nextflow (v.21.04.0) to analyse RNA-seq data[66]. Raw reads in fastq files were mapped to GRCm[38] with associated ensemble transcript definitions using STAR (v.2.7.6a)[67]. BAM files were sorted with a chromosome coordinate using samtools (v.1.12). RSEM (v.1.3.1) was used to calculate estimated read

counts per gene and to quantify a measure of TPM[68]. Differential expression analysis was performed using the R platform (v.4.0.3) package DESeq2 (ref. 69), filtering with the absolute value of log(fold change) > 1 and FDR < 0.05. Significantly differentially expressed genes were determined using a generalized linear model within DESeq2 and a Wald test. Gene expression levels between treatment groups was further analysed for their pathway enrichments using gene set enrichment analysis[70]. Normalization (using z-scores) of TPM scores across the dataset was performed before plotting heatmaps of gene expression. The AT2 activated score was derived using a previously described method[71]. In brief, bulk RNA-seq data from mouse models, with or without an EGFR mutation and in the presence or absence of PM exposure, were compared according to the degree to which they were similar to a signature of activated AT2 transitional progenitor cells (‘AT2 activated’) derived from previously published single-cell RNA-seq data[23]. This signature was estimated using a pseudoR2 value calculated using a previously described approach[71]. This approach was adapted to a mouse dataset using gene weights from mouse-to-human orthologous genes. The pseudoR2 value was used as a continuous input in a test between the different conditions. Comparison of RNA-seq data from mice to never-smokers in the COPA study. RNA-seq was applied to 18 samples of bronchial brushings from nine never-smokers from the COPA study after exposure to filtered air and diesel exhaust. Salmon[72] was used to estimate transcript-level abundance from RNA-seq read data. Differential expression analysis was performed using DESeq2 (ref. 69). The log twofold difference in gene expression was calculated between samples collected 24 h after exposure to diesel exhaust and filtered air (control) on separate occasions but from the same participants. P values were adjusted using the Benjamini–Hochberg method. The log twofold change of significantly differentially expressed genes between the tdTomato control and td-Tomato PM-treated mice were compared to the log twofold change expression of the genes from COPA participants. All participants in the COPA study provided informed consent. The consent forms and study protocol were approved by the University of British Columbia Clinical Research Ethics Board (number H12-03025), Vancouver Coastal Health Ethics Board (number V12-03025) and Health Canada's Research Ethics Board (number 2012–0040). The limitation of this analysis is that the mouse and human RNA-seq datasets fundamentally differ in the following ways. (1) Mouse data were acquired from total lung $EpCAM^+$ cells, containing both airway and alveolar tissue, whereas the human data were obtained from bronchial brushings only; therefore, different cell types are represented in the data. (2) The pollution exposure between species differed. Human participants were exposed to diesel exhaust for 2 h compared to 3 weeks of PM exposure for mice. Furthermore, the mice were kept in controlled environments, whereas a 4-week washout period between exposure to filtered air and diesel exhaust in human participants was required, where day-to-day PM exposures and lifestyle differences could not be controlled. (3) Fold changes from the human data were obtained by pairwise comparisons from each individual. By contrast, because we did not have pairwise matched data from each mouse, the fold changes from the mouse data were derived based on aggregated (mean) values across each condition (that is, air pollution versus control). (4) The RNA-seq was performed at two different sequencing centres and target depths were different. The human data were sequenced with a target depth of 30 million reads per sample, whereas the mouse data were sequenced with a target depth of 25 million reads per sample.

Organoid-forming assays

Lung organoid co-culture assays have been previously described[22]. In brief, tdTomato$^+$ lung epithelial cells (tdTomato$^+$EpCAM$^+$CD45$^-$CD31$^-$TER119$^-$) and tdTomato$^-$ lung epithelial cells (tdTomato$^-$EpCAM$^+$CD45$^-$CD31$^-$TER119$^-$) were isolated by FACS from PBS-treated or PM-exposed ET mice after 3 weeks of treatment and were resuspended in 3D organoid medium consisting of DMEM/F12 with 10% FBS, 100 U ml^{-1} penicillin–streptomycin, insulin, transferrin, selenium, l-glutamine (all from Gibco) and 1 mM HEPES (in-house). About 5,000–10,000 cells were mixed with a mouse lung fibroblast cell line (MLg2908, American Type Culture Collection, 1:5 ratio) and resuspended in growth-factor-reduced Matrigel (Corning) at a ratio of 1:1. Next 100 µl of this mixture was pipetted into a 24-well Transwell insert with a 0.4 µm pore (Corning). After incubating for 30 min at 37 µC, 500 µl of organoid medium was added to the lower chamber and the medium changed every other day. Bright-field and fluorescence images were acquired after 14 days using an EVOS microscope (Thermo Fisher Scientific) and quantified using Fiji (v.2.0.0-rc-69/1.52r; ImageJ). For ex vivo IL-1β treatment of lung AT2 cells, single-cell suspensions from ET mice lungs (without in vivo Cre induction) were subject to AT2 cell purification as previously described (MHC Class II$^+$CD49f^{low} EpCAM$^+$CD45$^-$CD31$^-$TER119$^-$)[61]. Purified AT2 cells were incubated in vitro with 6 x 10^7 p.f.u. ml^{-1} of Ad5-CMV-Cre in 100 pl per 100,000 cells in 3D organoid medium for 1 h at 37 µC as previously detailed[27]. Cells were washed three times in PBS before plating as described above, and 20 ng ml^{-1} IL-1β was added to the organoid medium in the lower chamber and changed every other day. TdTomato$^+$ organoids were quantified in Fiji. For whole-mount staining of organoids, organoids were prepared according to previous published methods[73] and stained with anti-proSPC (Abcam, clone EPR19839) and anti-KRT8 (DSHB Iowa, clone TROMA-1). 3D confocal images were acquired using an Olympus FV3000 microscope and analysed using Fiji. For assessment of AT2 organoid formation after PM exposure, AT2 cells were isolated from PBS-treated or PM-treated control mice and ET mice after 3 weeks, without in vivo Cre induction. Following Cre infection, 10,000 cells were plated in the organoid assay as described above. For co-culture of AT2 cells and macrophages, non-induced ET mice were exposed to either PBS or PM, followed by collection at 3 weeks, and AT2 cells, interstitial cells and alveolar macrophages were isolated as previously detailed[22] (sorting strategies are defined in Extended Data Fig. 6c). AT2 cells from PBS-treated ET mice only were infected with Cre ex vivo as described above, before 10,000 AT2 cells were either plated with fibroblasts only or with a 1:6 ratio of PBS-treated or PM-treated macrophages as described above, modified from a previously published method[22]. tdTomato$^+$ organoids were quantified in all conditions.

Statistics and reproducibility

Preclinical statistical analyses were performed using Prism (v.9.1.1, GraphPad Software) with centre line depicting median unless otherwise stated. Analyses of epidemiological data and mutation and sequence data were performed in R (v.3.6.2. or v.4.1.3 (UKBB analysis)). Graphic display was performed in Prism, and illustrations in Fig. 3c,e and Extended Data Figs. 1a, 5d,f and 8a were created using BioRender (https://biorender.com). A Kolmogorov–Smirnov normality test was performed before any other statistical test.

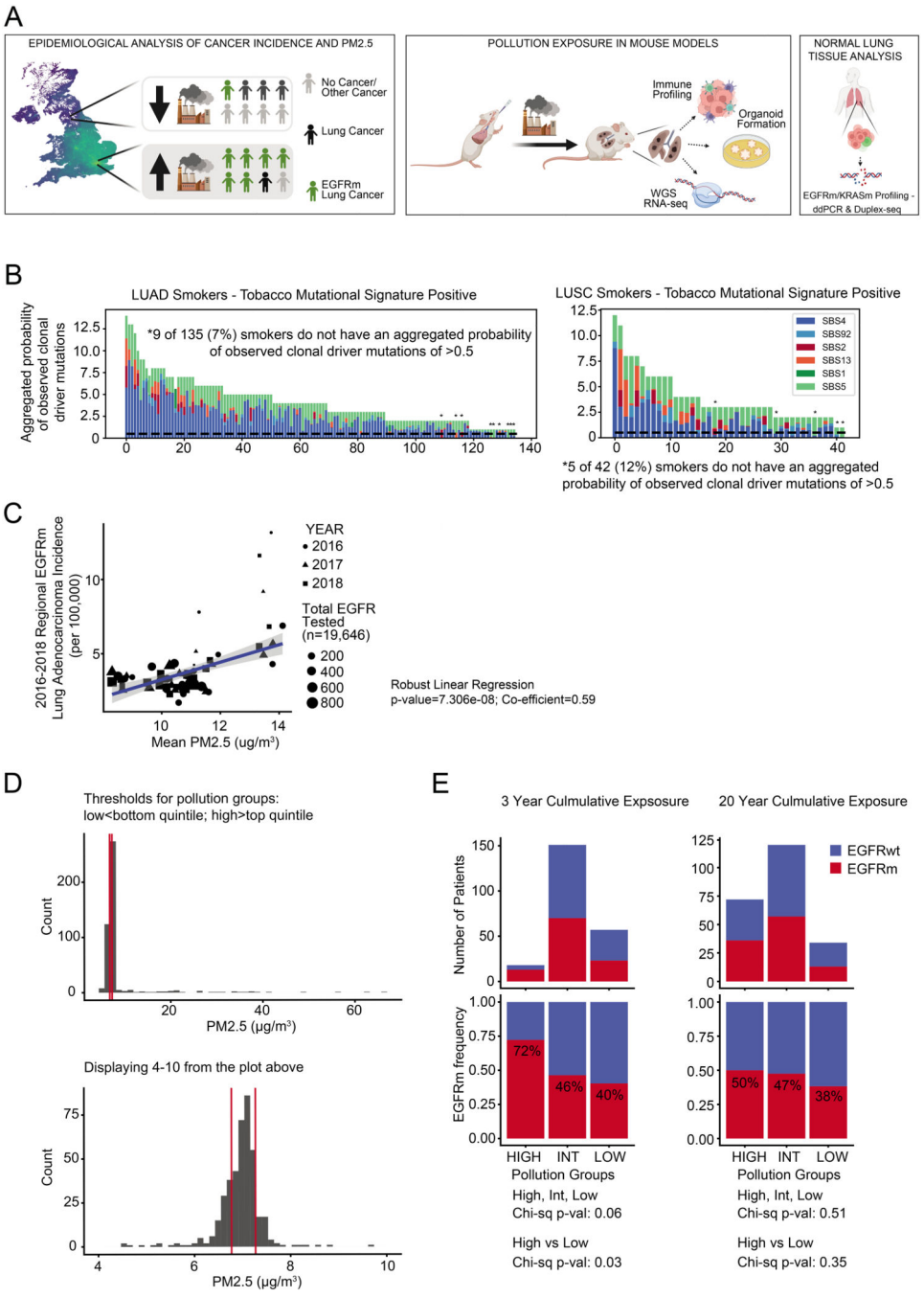
Europe PMC Funders Author Manuscripts

Afterwards, if any of the comparative groups failed normality (or the number was too low to estimate normality), a nonparametric Mann–Whitney test was performed. When groups showed a normal distribution, an unpaired two-tailed t-test was performed. When groups showed a significant difference in the variance, we used a t-test with Welch's correction. When assessing statistics of three or more groups, we performed ANOVA or nonparametric Kruskal–Wallis test controlling for multiple comparisons. Blinded analysis was carried out for all image and tumour analysis. No data were excluded. No statistical methods were used to predetermine sample sizes in the mouse studies. Mice with matched sex and age were randomized into different treatment groups. All experiments were reliably reproduced. Specifically, all in vivo experiments, except for omics data (RNA-seq), were performed independently at least twice, with the total number of biological replicates (independent mice) indicated in the corresponding figure legends.

Driver mutation probability

The list of driver mutations and the mutational signature exposures were obtained from the TRACERx 421 publication[14]. Only patients with detected smoking-related signatures are considered in the analysis (TRACERx 421). Each observed clonal driver mutation was given a probability to be caused by all active mutational signatures in the patient. This number was derived by multiplying the exposures of the mutational signatures with the 96-channel profile of each signature[74]. Then the value was normalized to 1 so that each driver mutation can be explained by a fraction of active mutational signatures. The probabilities were then aggregated, giving the overall contribution to driver mutations from each of the active mutational signatures. A patient was defined as non-carrier of a tobacco-related driver mutation if the probability of SBS4 and SBS92 (smoking-related signatures) was less than 0.5.

Extended Data

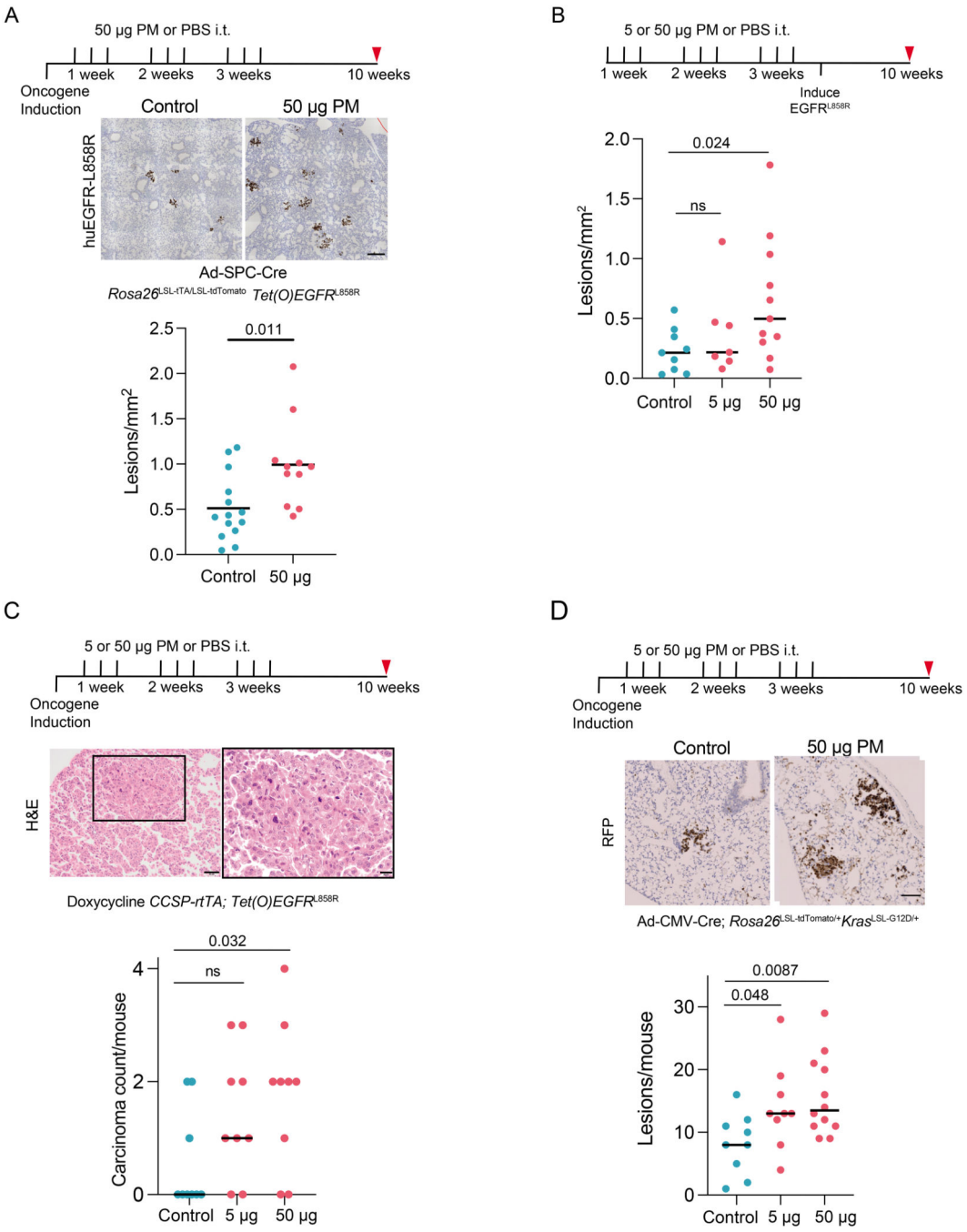


Extended Data Figure 1.

A) Study design schematic featuring the 3 aspects of the paper. LEFT: Epidemiological analysis of cancer incidence and $PM_{2.5}$. MIDDLE: Pollution exposure in mouse models. RIGHT: Normal lung tissue analysis. B) TX421 Tumours from Smokers. Barplots indicating proportion of SNVs in each tumour attributed to each SBS mutational signature. The barplots (Top: Lung adenocarcinoma (LUAD), Bottom: Lung sqaumous cell carcinoma (LUSC)) reflect the probability that clonal driver mutations in patients, where smoking-

related signatures have been detected, are caused by different mutational processes (SBS4 and SBS92 smoking, SBS2 and SBS13 APOBEC, SBS1 and SBS5 ageing). Each observed driver mutation in each patient is given a mutational-signature-causing probability based on the trinucleotide context and the signatures exposure of the patient (see Methods) and then these probabilities are aggregated. Asterisks represent patients where the smoking-related aggregated probabilities are below 0.5. C) Correlation between $PM_{2.5}$ levels and *EGFR* mutant (EGFRm) adenocarcinoma lung cancer incidence in England. The blue line: robust linear regression line; grey shading: 95% confidence interval. D-E) The Canadian Lung Cancer Cohort. D) Distribution of 3 year and 20 year cumulative $PM_{2.5}$ exposure levels for all patients in the Canadian cohort. Red lines mark the thresholds that were used to determine Low, Intermediate and High groups that are used in (D). These are the 1st (6.77ug/m^3) and 5th quintiles (7.27ug/m^3) of the distribution. The full distribution is displayed in the top plot, while the bottom plot displays a narrower range of 4-10 ug/m^3 (for clarity). E) Counts and frequencies of EGFRm in the Canadian Cohort, where 3 year and 20 year cumulative $PM_{2.5}$ exposure levels were available. Patients are grouped into high, intermediate and low groups based on thresholds established as described in (D). These groups are defined based on 3 year cumulative $PM_{2.5}$ exposure data (left) and based on 20 year cumulative $PM_{2.5}$ exposure data (right). The bar plots display the counts and frequency of EGFRm amongst patients within each group

Europe PMC Funders Author Manuscripts

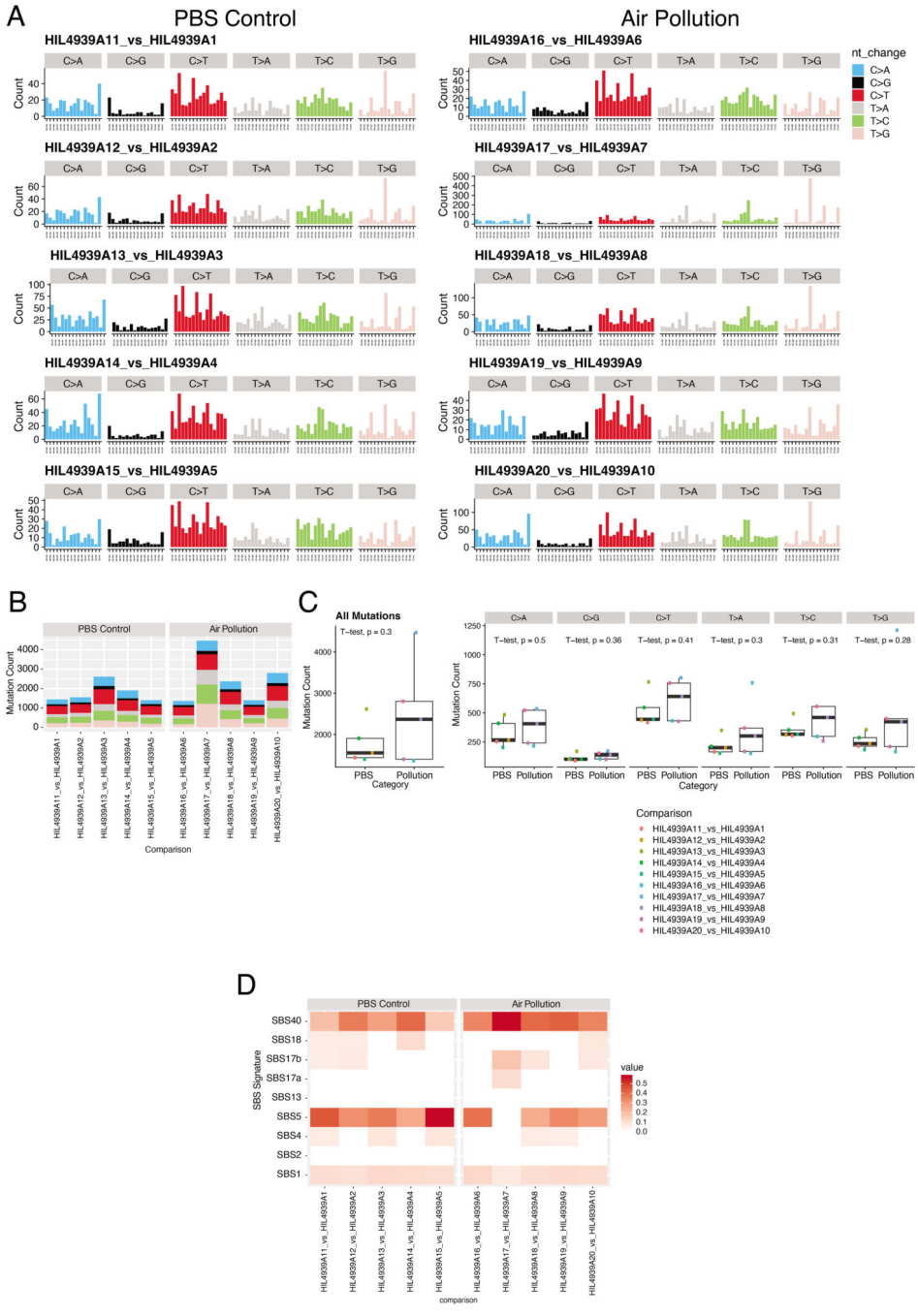


Extended Data Figure 2.

A) Schematic of PM exposure and representative IHC of ET mice induced with AT2-specific Ad5-SPC-Cre exposed to PM or PBS control and quantification of neoplastic lesions (n=14 PBS, n=11 PM). Mann-Whitney test. B) Schematic of PM exposure followed by expression of *EGFR*^{L858R} and quantification of precancerous lesions/mm^2 of lung tissue (n=9 PBS; n=7 5 µg; n=11 50 µg PM). One-way ANOVA. C) Schematic of PM exposure and representative H&E of a lung adenocarcinoma in a 50 µg PM exposed, doxycycline treated *CCSP-rtTa; TetO-EGFR*^{858R} mice; quantification of number

Europe PMC Funders Author Manuscripts

of adenocarcinomas per mouse below (n = 9 per group). One-way ANOVA. D) Schematic of PM exposure and representative IHC for red fluorescent protein (RFP, marks tdTomato+ cells) in *Rosa26*$^{\text{LSL-tdTomato/+}}$;*Kras*$^{\text{LSL-G12D/+}}$ mouse model in control or 50 μg PM exposed conditions; quantification of number of hyperplastic lesions per mouse (n= 9 control, n=9 5 μg and n=12 50 μg). One-way ANOVA. Scale bar 50 μm (C main, H), 20 μm (C insert), 100 μm A & D

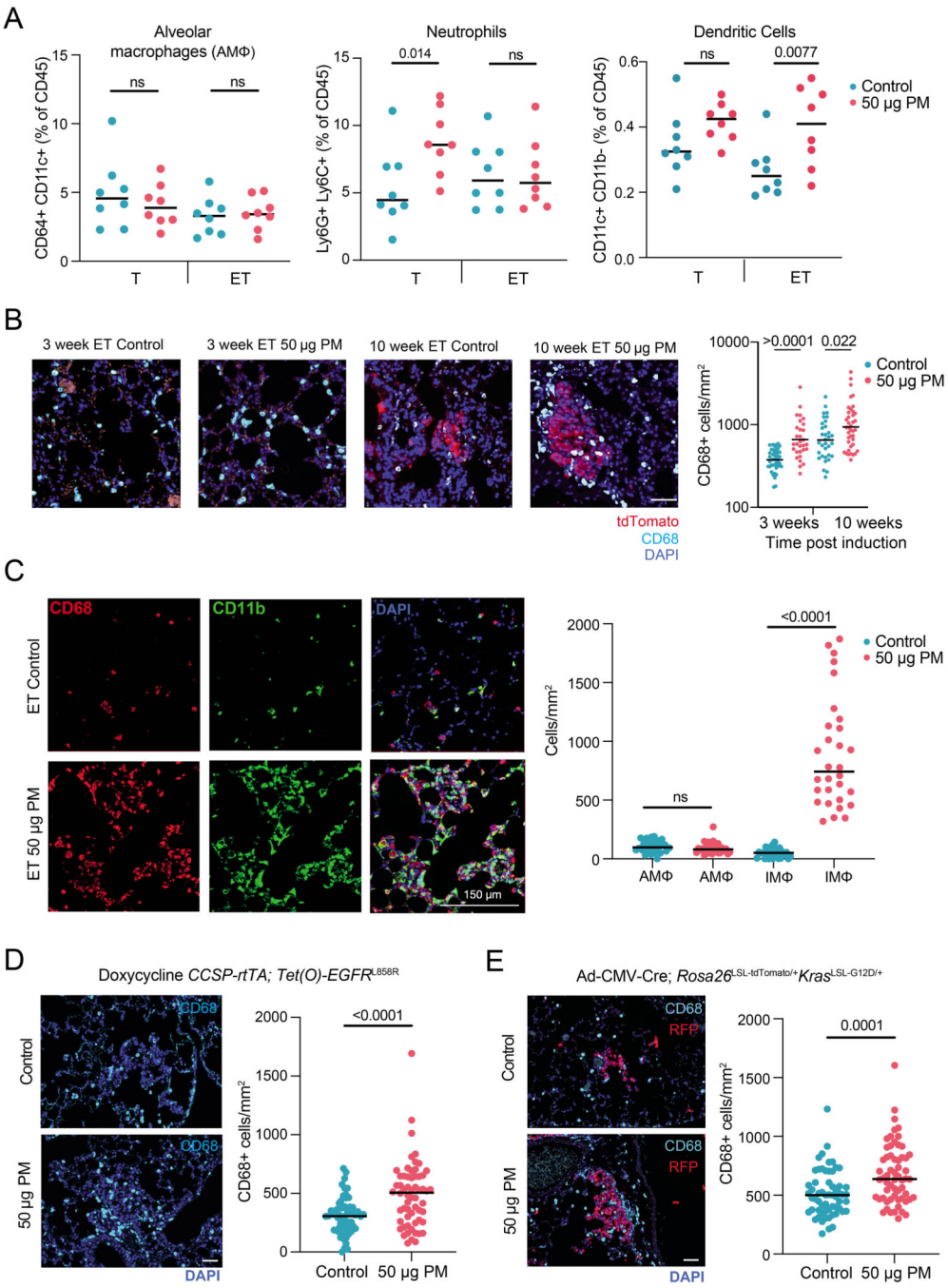


Extended Data Figure 3.

Europe PMC Funders Author Manuscripts

WGS analysis of tumours from ET mice exposed to air pollution (n=5) and those exposed to PBS controls (n=5). Each mouse tumour is compared vs the corresponding germline from the same mouse. A) Mutational profiles for each tumour sample according to the mutation trinucleotide context. LEFT: PBS Controls, RIGHT: 50 μg PM. B) Barplots indicate the counts of mutations in each sample, where bars are colored based on the base change. C) Boxplot comparing the counts of mutations between tumours from pollution exposed mice (50 μg PM) and tumours from PBS exposed mice (PBS Control). All mutations are summarised in one plot on the left, and are then further divided based on the base change of the mutation (n=5 mice per group). Two-sided T-test comparing numbers of mutations between PBS and Air Pollution p-values are displayed. The boxplot line represents the median, the hinges of the box represent the 1st and 3rd quartiles and the limits of the whiskers represent the 1.5 interquartile range. D) Attribution of mutations in each tumour sample to each single base substitution (SBS) mutation signature. The shading indicates the weight of the signature within each sample. Majority of the weights have been assigned to ageing related signatures (SBS40, SBS5, SBS1) Komogolomov-Smirnoff test p-value=0.26-0.68

Europe PMC Funders Author Manuscripts

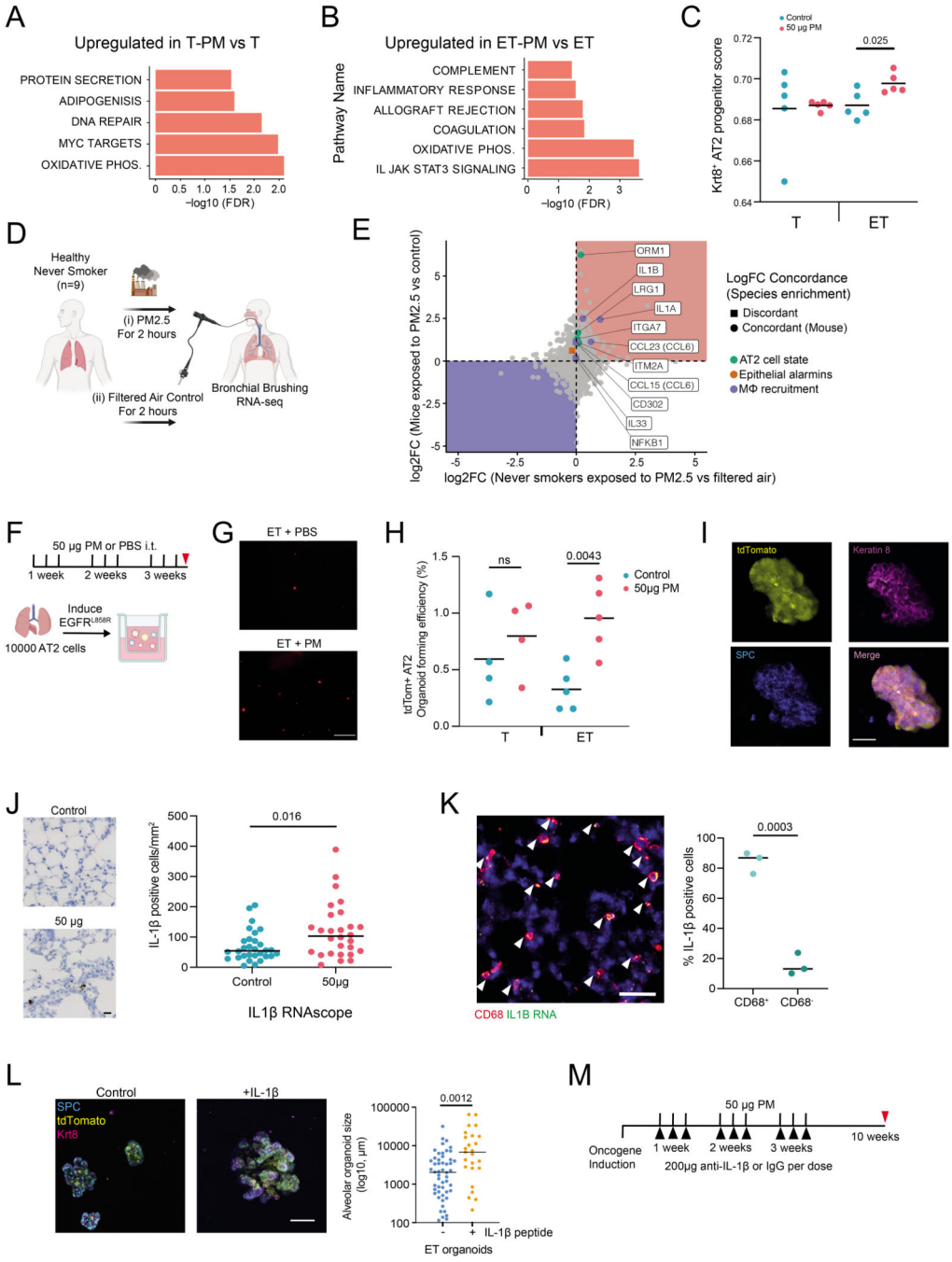


Extended Data Figure 4.

A) Immune cell frequencies in the lungs determined by flow cytometry 24 hours post-exposure from induced tdTomato (T) and *EGFR* mutant (ET) mice after 50 µg PM (red) or control (blue) (n=8 mice per group). Data are presented as the frequency among live CD45+ immune cells. One-way ANOVA. B) Representative immunofluorescent images of CD68+ macrophages (cyan) and tdTomato+ *EGFR* mutant cells (red) within ET lungs exposed to control or 50 µg PM. Quantification of CD68+ cells per mm^2 of lung tissue (n= 4 mice per group). C) Representative immunofluorescent images of CD68 (red), CD11b

Europe PMC Funders Author Manuscripts

(green) and merged images from induced ET mice after 3 weeks of exposure to PBS (top) or 50 μg PM (bottom). Quantification of alveolar macrophages (AMΦ, CD68+CD11b-) and interstitial macrophages (IMΦ, CD68+CD11b+) per mm^2 of lung tissue, selecting 10 x random 500 μm^2 fields of view per mouse (n= 3 mice per group). One-way ANOVA. D) Representative immunofluorescent images of CD68+ macrophages (cyan) within *CCSP-rtTA; TetO-EGFR*^{L858R} lungs treated with PBS (top) or 50 μg PM (bottom) 10 weeks post oncogene induction; quantification of CD68+ cells per m^2 of lung tissue, selecting 20 x random 500 μm^2 fields of view per mouse (n= 3 mice per group). Unpaired t-test. E) Representative immunofluorescent images of CD68+ macrophages (cyan) and tdTomato+ *Kras*^{G12D} mutant cells (red) within KT lungs treated with PBS (top panel) or 50 μg PM (bottom) 10 weeks post oncogene induction; quantification of CD68+ cells per mm^2 of lung tissue, selecting 20 x 500 μm^2 fields of view containing RFP+ cells per mouse (n= 3 mice per group). Unpaired t-test. Scale bar 50 μm B & D, 150 μm E. Gating strategies for flow cytometry analysis provided in Extended Data Figure 6. Statistical analysis by one-way ANOVA for B, D, E & G. Scale bars 100 μm (B,F,E)



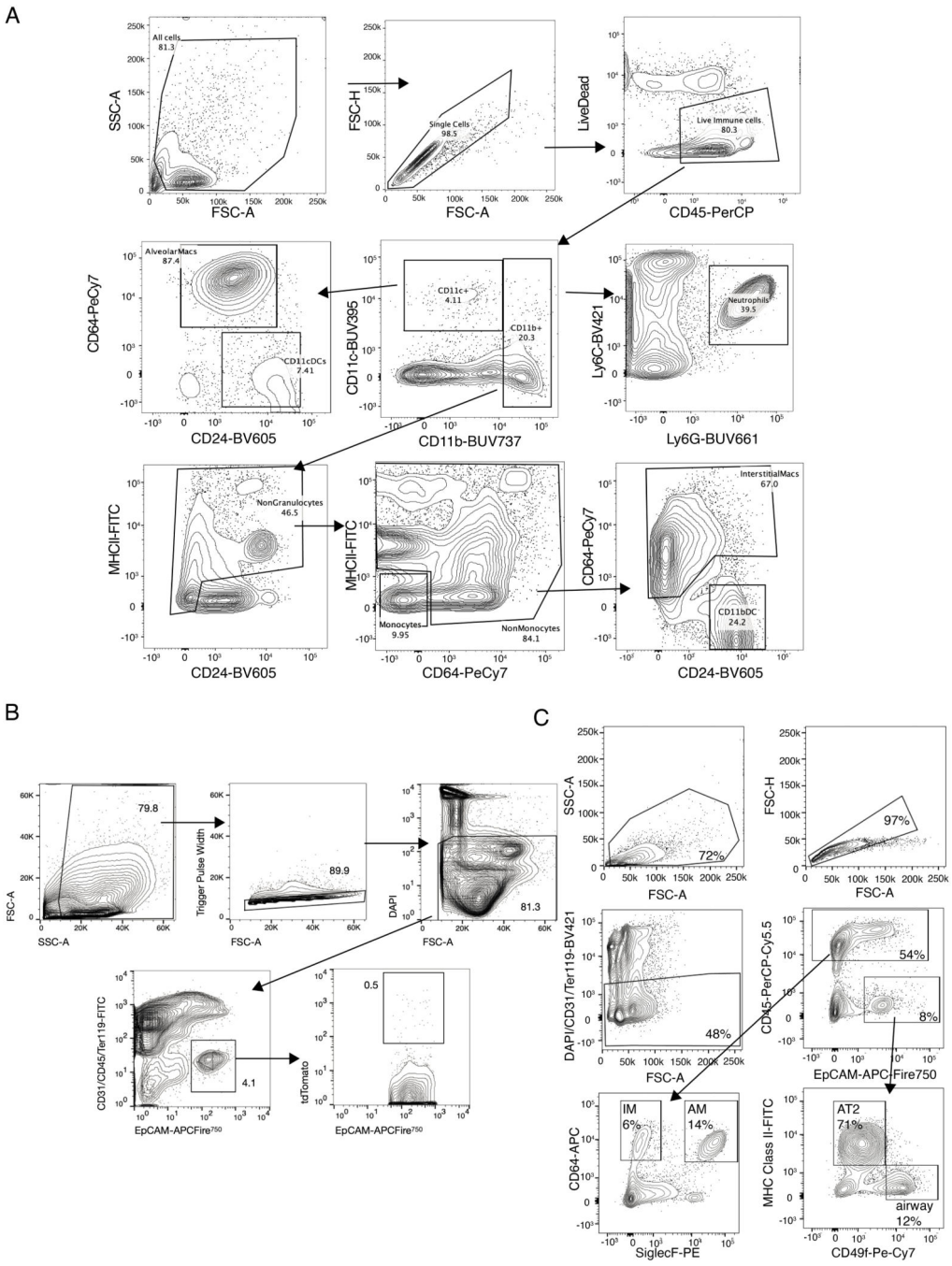
Extended Data Figure 5.

A-B) Significantly enriched GSEA pathways upregulated in T-PM lung epithelial cells compared to T control mice (A), in ET-PM lung epithelial cells compared to ET control mice (B). For each comparison, barplots indicate the -log10(FDR) of the Komogolomov-Smirnoff test p-value for each pathway. C) Krt8+ AT2 progenitor score derived from scRNAseq of bleomycin treated mouse lung used to deconvolute bulk RNA-seq of T and ET mice exposed to 50 µg PM or PBS, (n= 5 mice per group). Welch's t-test between control and PM. D) Schematic displaying experimental set-up of clinical exposure study in

never-smoker volunteers, crossover design with (i) and (ii) in random order separated by 4-week washout. E) Fold change (FC) of significantly upregulated genes (identified in mouse) compared to the fold change of genes changed in the clinical exposure study. Common directionality across species indicated by colour (negative: blue background; positive: red background). F) Schematic of AT2 culture from T or ET mice exposed to 50 μg PM or PBS, with induction of tdTomato or oncogene *ex vivo*. G) Representative fluorescent images of tdTomato+ AT2 organoids at day 14 from ET mice exposed to PBS or 50 μg PM in vivo. Scale bar 100 μm. H) Quantification of tdTom+ AT2 organoid forming efficiency, data represents averages from 2 technical replicates/mouse; n=4 mice from T control and PM; n=5 mice for ET control and PM. One-way ANOVA. I) Representative fluorescent imaging of tdTomato (yellow), Keratin 8 (magenta), SPC (blue) on a wholemount AT2 organoid from an ET mouse treated with 50 μg PM. Scale bar is 20 μm. J) LEFT: Representative IL-1β RNAscope performed on lungs from ET mice treated with PBS or 50 μg PM after 3 weeks of exposure. Scale bar 20 μm. RIGHT: Quantification of IL-1β+ cells per mm^2 of lung tissue from 30 random fields of view (control, n = 3 mice) and 28 fields of view (50 μg PM, n = 3 mice). Mann-Whitney test p-value is displayed. K) LEFT: Representative image of IL-1β RNAscope (green) in CD68 positive (red) macrophages, arrows indicate positive macrophages. n=3 mice per group. Scale bar 50 μm. RIGHT: Quantification of IL-1β positive CD68+ cells at 3 weeks post induction in ET mice following exposure to PM. Mann-Whitney test. L) LEFT: Representative fluorescent images of *EGFR*^{L858R} naive (non-PM exposed) AT2 organoids from ET mice treated with control or IL-1β *in vitro*. tdTomato (yellow) organoids stained with SPC (blue) and Keratin 8 (magenta). Scale bar 50 μm. RIGHT: Quantification of organoid size with each dot representing an organoid at day 14 of control (blue) or IL-1β treated (orange). Organoids derived from n=2 mice per group. Mann-Whitney test. M) Schematic of anti-IL-1β treatment treatment (black triangles) during PM exposure (black lines) and harvest (red triangle)

Europe PMC Funders Author Manuscripts

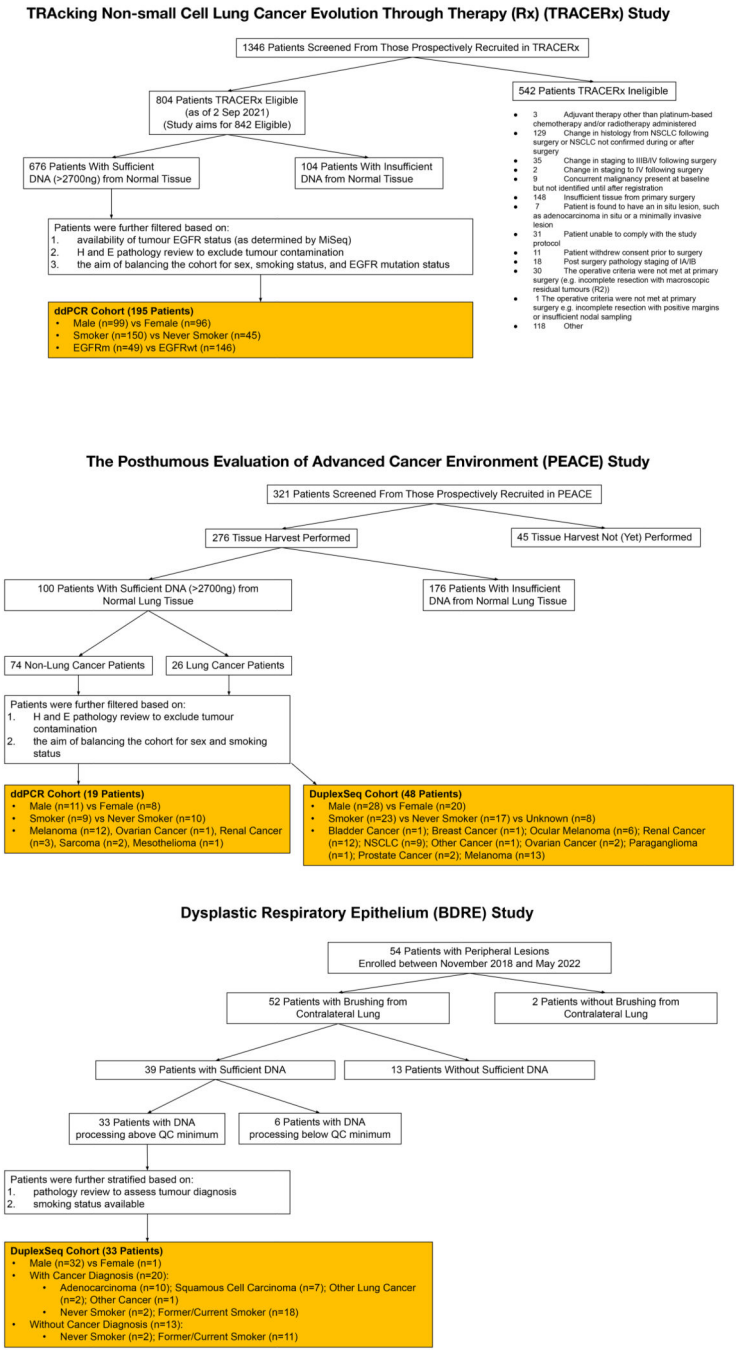
Europe PMC Funders Author Manuscripts



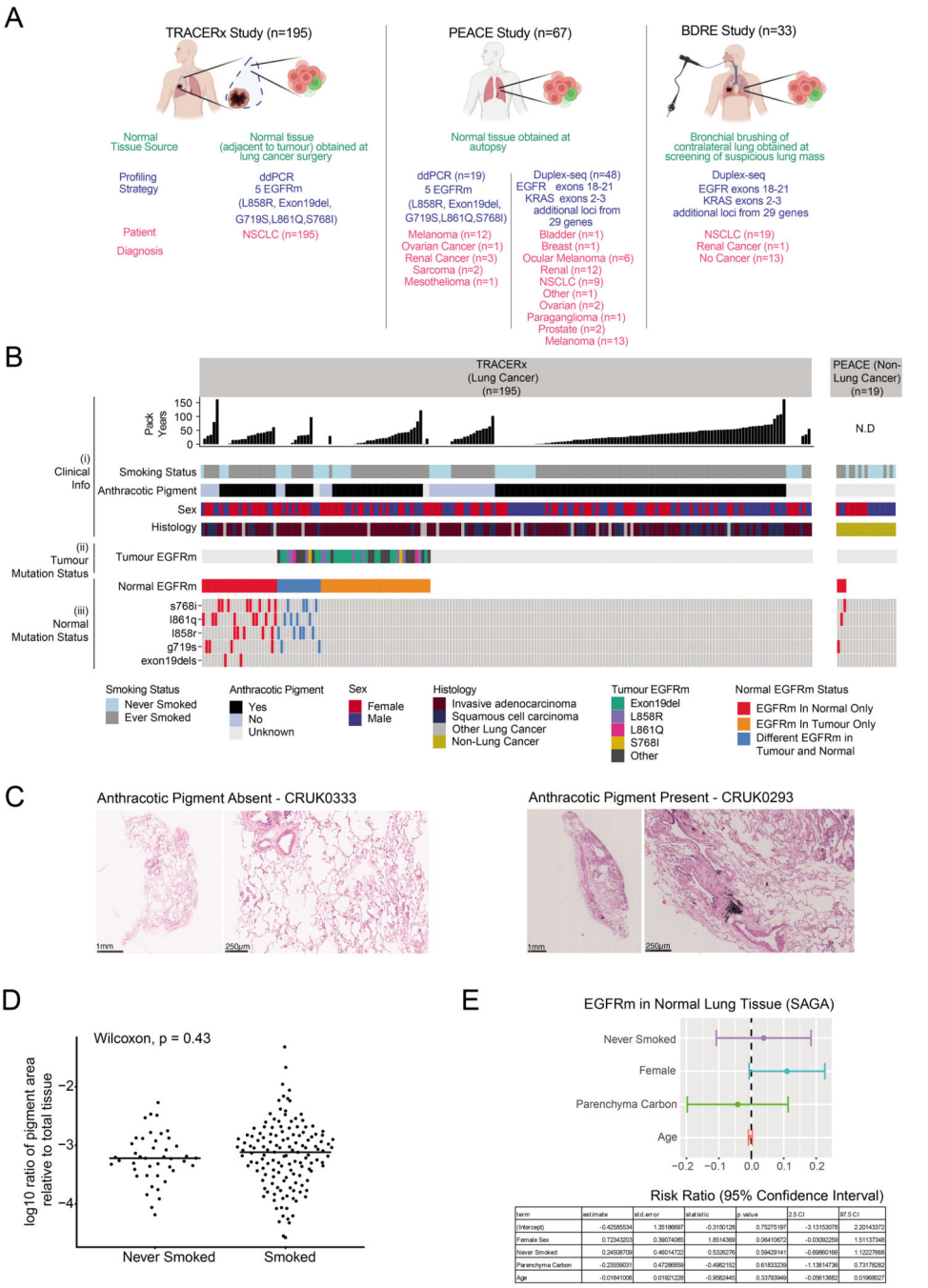
Extended Data Figure 6.

A, B) Example of flow gating strategy to determine frequency of lung (A) alveolar macrophages, interstitial macrophages, neutrophils, dendritic cells and (B) epithelial cells both tdTomato positive and negative. All samples were first gated to exclude debris and doublets, followed by live cell discrimination. C) Representative picture from a tdTomato mouse treated with control PBS for 3 weeks using sort strategy to enrich for for AT2 cells defined in Major et al., 2020 and both alveolar and interstitial macrophages defined in Choi et al., 2020

Europe PMC Funders Author Manuscripts



Extended Data Figure 7. CONSORT Diagrams for the normal lung tissue profiling cohorts

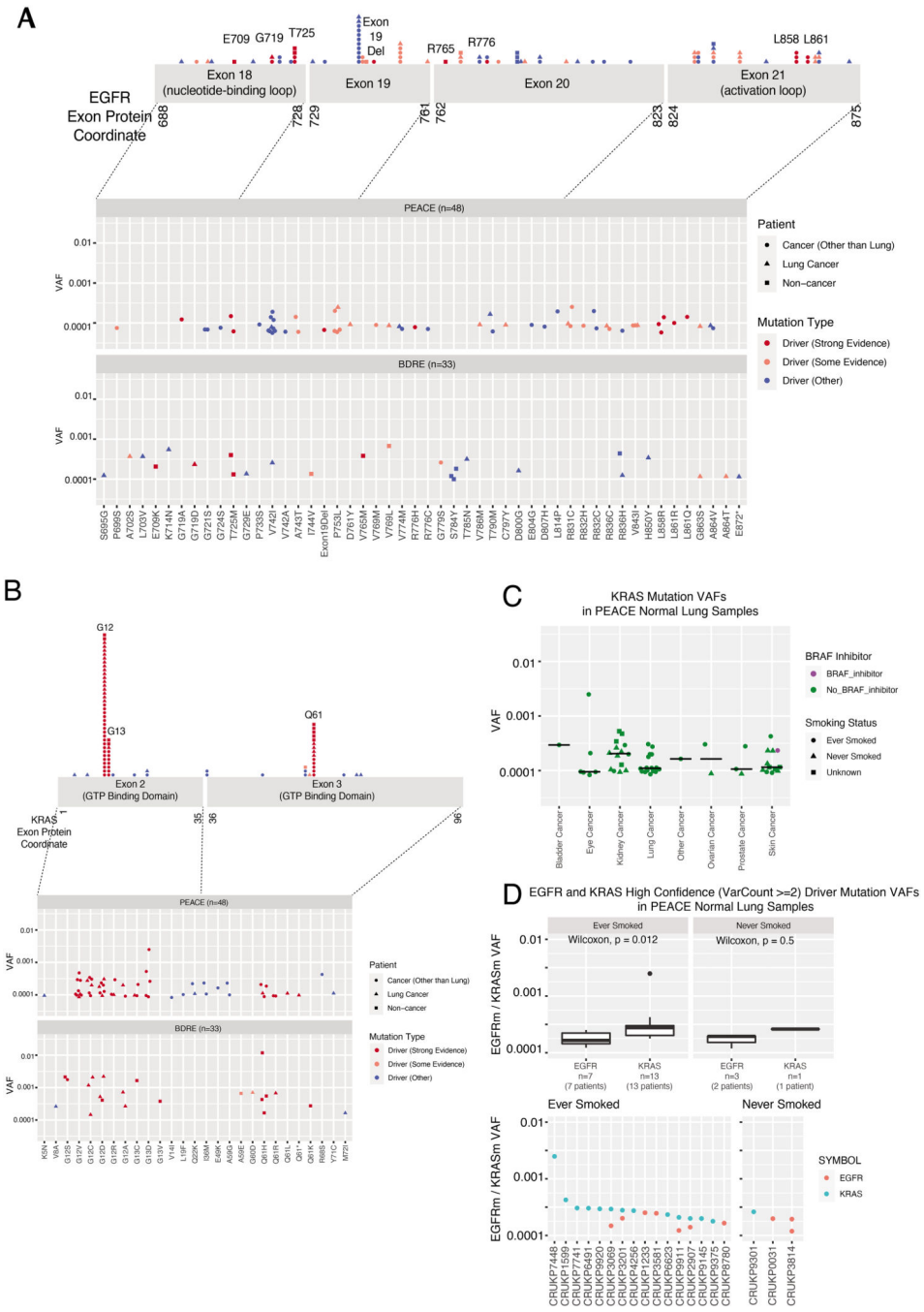


Extended Data Figure 8.

A) Schematic indicating normal lung tissue cohorts analysed by ddPCR and Duplex-seq. B) TRACERx and PEACE Cohort for ddPCR of 5 *EGFR* mutations. (i) Clinical information for each patient, (ii) Tumour *EGFR* mutation status, (iii) Normal *EGFR* mutation status. C) Representative H & E images from anthracotic pigment identification in TRACERx normal tissue. D) Comparing area of normal tissue harbouring anthracotic pigment in never smokers (n=43) and smokers (n=138). Each dot represents the ratio of pigmented area respective to total tissue in each anthracosis positive normal lung tissue sample. Two-sided Wilcox

Europe PMC Funders Author Manuscripts

test p-value is reported. E) Regression analysis of characteristics influences *EGFR* mutant (EGFRm) presence in normal lung tissue for ddPCR-TRACERx cohort (n=195)



Extended Data Figure 9.

A) Top: *EGFR* Mutations detected using Duplex-seq across *EGFR* exons 18-21 on normal lung samples from the BDRE Study. Bottom: VAFs of each *EGFR* mutation are displayed. B) Top: *KRAS* Mutations detected using Duplex-seq across *KRAS* exons 2-3 on normal lung samples from the BDRE Study. Bottom: VAFs of each *KRAS* mutation are displayed.

A-B) Only cancer-related mutations annotated in the cancer gene census are displayed. Mutations with strong evidence of being a lung cancer driver mutation are indicated in red, while mutations with some evidence of being a lung cancer driver mutation are indicated in pink, all other drivers annotated in COSMIC are indicated in blue. C) VAFs of *KRAS* mutations across samples of different cancer types. The one patient who received BRAF inhibitor treatment is indicated in purple. D) Comparing VAFs of high confidence (var count >=2, strong evidence) driver mutations in *EGFR* and *KRAS*. TOP: Boxplots summarise VAFs across samples. The boxplot line represents the median, the hinges of the box represent the 1st and 3rd quartiles and the limits of the whiskers represent the 1.5 interquartile range. Mutations are grouped according to the gene harbouring the mutation and smoking status of the patient. Two-sided Wilcox test p-values are reported. BOTTOM: dot plots show VAFs of mutations in each sample. Where a sample has 2 mutations (n=4), they are both indicated. Dots are coloured by the gene harbouring the mutation (*EGFR* or *KRAS*). A paired t-test was performed between the VAFs of *EGFR* and *KRAS* mutations in these 4 cases. (Paired t-test p=0.015) (Details of driver mutations can be found in Supplementary Table S8)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

William Hill[#1], Emilia L. Lim[#1,2,127], Clare E. Weeden[#1], Claudia Lee[1,2,3], Marcellus Augustine[1,2,3,4], Kezhong Chen[2,5], Feng-Che Kuan[6,7], Fabio Marongiu[8,9], Edward J. Evans Jr[8], David A. Moore[1,2,10], Felipe S. Rodrigues[11], Oriol Pich[1], Bjorn Bakker[1], Hongui Cha[2,12], Renelle Myers[13], Febe van Maldegem[14,15], Jesse Boumelha[14], Selvaraju Veeriah[2], Andrew Rowan[1], Cristina Naceur-Lombardelli[2], Takahiro Karasaki[1,2,16], Monica Sivakumar[2], Swapnanil De[2], Deborah R. Caswell[1], Ai Nagano[1,2], James R. M. Black[2,17], Carlos Martínez-Ruiz[2,17], Min Hyung Ryu[18], Ryan D. Huff[18], 1 Shijia Li[8], Marie-Julie Favé[19], Alastair Magness[1,2], Alejandro Suárez-Bonnet[20,21], Simon L. Priestnall[20,21], Margreet Lüchtenborg[22,23], Katrina Lavelle[22], Joanna Pethick[22], Steven Hardy[22], Fiona E. McRonald[22], Meng-Hung Lin[24], Clara I. Troccoli[8,25], Moumita Ghosh[26], York E. Miller[26,27], Daniel T. Merrick[28], Robert L. Keith[26,27], Maise Al Bakir[1,2], Chris Bailey[1], Mark S. Hill[1], Lao H. Saal[29,30], Yilun Chen[29,30], Anthony M. George[29,30], Christopher Abbosh[2], Nnennaya Kanu[2], Se-Hoon Lee[12], Nicholas McGranahan[2,17], Christine D. Berg[31], Peter Sasieni[32], Richard Houlston[33], Clare Turnbull[33], Stephen Lam[13], Philip Awadalla[19], Eva Grönroos[1], Julian Downward[14], Tyler Jacks[34,35], Christopher Carlsten[18], Ilaria Malanchi[11], Allan Hackshaw[36], Kevin Litchfield[2,4], TRACERx Consortium[*],

James DeGregori[8,127], Mariam Jamal-Hanjani[2,16,37,127], Charles Swanton[1,2,37,128]

Affiliations

[1]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK

[2]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK

[3]Division of Medicine, University College London, London, UK

[4]Tumour Immunogenomics and Immunosurveillance Laboratory, University College London Cancer Institute, London, UK

[5]Department of Thoracic Surgery and Thoracic Oncology Institute, Peking University People's Hospital, Beijing, China

[6]Department of Hematology and Oncology, Chang Gung Memorial Hospital, Chiayi Branch, Chiayi, Taiwan

[7]Graduate Institute of Clinical Medical Sciences, Chang-Gung University, Taoyuan, Taiwan

[8]Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[9]Department of Biomedical Sciences, University of Cagliari, Cagliari, Italy

[10]Department of Cellular Pathology, University College London Hospitals, London, UK

[11]Tumour-Host Interaction Laboratory, The Francis Crick Institute, London, UK

[12]Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

[13]BC Cancer Research Institute, University of British Columbia, Vancouver, British Columbia, Canada

[14]Oncogene Biology Laboratory, The Francis Crick Institute, London, UK

[15]Department of Molecular Cell Biology and Immunology, Amsterdam UMC, Amsterdam, The Netherlands

[16]Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK

[17]Cancer Genome Evolution Research Group, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK

[18]Department of Medicine, Division of Respiratory Medicine, Chan-Yeung Centre for Occupational and Environmental Respiratory Disease, Vancouver Coastal Health Research Institute, UBC, Vancouver, British Columbia, Canada

[19]Ontario Institute for Cancer Research, Toronto, Ontario, Canada

[20]Department of Pathobiology and Population Sciences, The Royal Veterinary College, Hatfield, UK

[21]Experimental Histopathology, The Francis Crick Institute, London, UK

[22]National Disease Registration Service (NDRS), NHS England, Leeds, UK

[23]Centre for Cancer, Society and Public Health, Comprehensive Cancer Centre, School of Cancer and Pharmaceutical Sciences, King's College London, London, UK

[24]Health Information and Epidemiology Laboratory, Chang-Gung Memorial Hospital, Chiayi, Taiwan

[25]Flagship Biosciences, Boulder, CO, USA

[26]Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[27]Veterans Affairs Eastern Colorado Healthcare System, Aurora, CO, USA

[28]Department of Pathology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[29]SAGA Diagnostics, Lund, Sweden

[30]Division of Oncology, Department of Clinical Sciences, Lund University, Lund, Sweden

[31]Early Cancer Detection Consultant, Bethesda, MD, USA

[32]Comprehensive Cancer Centre, King's College London, London, UK

[33]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK

[34]David H. Koch Institute for Integrative Cancer Research, Cambridge, MA, USA

[35]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

[36]Cancer Research UK and UCL Cancer Trials Centre, London, UK

[37]Department of Oncology, University College London Hospitals, London, UK

Acknowledgements

This research was conducted using the UK Biobank Resource under application number 82693. This work was supported by the Mark Foundation ASPIRE I Award (grant 21-029-ASP), the Lung Cancer Research Foundation Grant on Disparities in Lung Cancer, Advanced Grant (PROTEUS, grant agreement no. 835297), CRUK EDD (EDDPMA-Nov21\100034) and a Rosetrees Out-of-round Award (OoR2020\100009). W.H. is funded by an ERC Advanced Grant (PROTEUS, grant agreement no. 835297), CRUK EDD (EDDPMA-Nov21\100034), The Mark Foundation (grant 21-029-ASP) and has been supported by Rosetrees. E.L.L. receives funding from the NovoNordisk Foundation (ID 16584), The Mark Foundation (grant 21-029-ASP) and has been supported by Rosetrees. C.E.W. is supported by a RESPIRE4 fellowship from the European Respiratory Society and Marie-Sklodowska-Curie Actions. C.L. is supported by the Agency for Science, Technology & Research, Singapore and the Cancer Research UK City of London Centre and the City of London Centre Clinical Academic Training Programme. M.A. is supported by the City of London Centre Clinical Academic Training Programme (Year 3, SEBSTF-2021\100007). K.C. is supported by the Research Unit of Intelligence Diagnosis and Treatment in Early Non-small Cell Lung Cancer, the Chinese Academy of Medical Sciences (2021RU002), the National Natural Science Foundation of China (no. 82072566) and Peking University People's Hospital Research and Development Funds (RS2019-01). T.K. receives grant support from JSPS Overseas Research Fellowships Program (202060447). S.-H.L. is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (no. 2020R1A2C3006535), the National Cancer Center Grant (NCC1911269-3) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number HR20C0025). L.H.S. receives grant support from the Berta Kamprad Foundation, the Swedish Cancer Society and the Swedish Research Council. R.M. and S.L. acknowledge funding from the Terry Fox Research Institute. N.M. is a Sir Henry Dale Fellow,

jointly funded by the Wellcome Trust and the Royal Society (grant number 211179/Z/18/Z) and receives funding from Cancer Research UK, the Rosetrees and the NIHR BRC at University College London Hospitals and the CRUK University College London Experimental Cancer Medicine Centre. J. DeGregori, M.G., Y.E.M., D.T.M. and R.L.K. receive funding from the American Association for Cancer Research/Johnson&Johnson (18-90-52-DEGR), and J. DeGregori is supported by the Courtenay C. and Lucy Patten Davis Endowed Chair in Lung Cancer Research and a Merit Award from the Veteran's Administration (1 I01 BX004495). M.G., Y.E.M., D.T.M. and R.L.K. were supported by the National Cancer Institute (NCI) RO1 CA219893. E.J.E.J. was supported by a NCI Ruth L. Kirschstein National Research Service Award T32-CA190216 and the Blumenthal Fellowship from the Linda Crnic Institute for Down Syndrome. C.I.T. acknowledges funding from UC Anschutz (LHNC T32CA174648). The work at the University of Colorado was also supported by NCI Cancer Center Support Grant P30CA046934. K. Litchfield is funded by the UK Medical Research Council (MR/P014712/1 and MR/V033077/1), the Rosetrees Trust and the Cotswold Trust (A2437) and Cancer Research UK (C69256/A30194). M.J.-H. is a CRUK Career Establishment Awardee has received funding from Cancer Research UK, IASLC International Lung Cancer Foundation, the National Institute for Health Research, the Rosetrees Trust, UKI NETs and the NIHR University College London Hospitals Biomedical Research Centre. C.S. is a Royal Society Napier Research Professor (RSRPÄRÄ210001). His work is supported by the Francis Crick Institute that receives its core funding from Cancer Research UK (CC2041), the UK Medical Research Council (CC2041), and the Wellcome Trust (CC2041). For the purpose of Open Access, the author has applied a CC BY public copyright licence 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); Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025); the Rosetrees Trust, Butterfield and Stoneygate Trusts; NovoNordisk Foundation (ID16584); Royal Society Professorship Enhancement Award (RP/EA/180007); National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre; the Cancer Research UK-University College London Centre; Experimental Cancer Medicine Centre; the Breast Cancer Research Foundation (US) (BCRF-22-157); Cancer Research UK Early Detection an Diagnosis Primer Award (grant EDDPMA-Nov21/100034); and The Mark Foundation for Cancer Research Aspire Award (grant 21-029-ASP). This work was supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (grant number: SU2C-AACR-DT23-17 to S.M. Dubinett and A.E. Spira). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Researchgrants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. C.S. is in receipt of 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). We acknowledge the PEACE Consortium (PEACE Consortium members are named below) for their expertise and support in putting together the healthy tissue sample cohorts. We thank the clinical and administrative team of the PEACE study for their assistance in data curation (S. Shepherd, Z. Tippu, B. Shum, C. Lewis, M. O'Flaherty, A. Lucanas, E. Carlyle, L. Holt, F. Williams); nursing and biospecimen coordinators for their assistance in sample curation (K. Edmonds, L. Grostate, K. Lingard, D. Kelly, J. Korteweg, L. Terry, J. Biano, A. Murra, K. Kelly, K. Peat, N. Hunter); A. H. -K. Cheung for assistance in pathology review; J. Asklin and C. Forsberg for logistical and technical assistance; staff at the Chang Gung Memorial Hospital for providing Chang Gung Research Database (CGRD) data; staff who provided support at the Flow Cytometry Unit, the Experimental Histopathology Unit, the Advanced Light Microscopy Facility, the Advanced Sequencing Facility and the Biological Resources Unit, especially N. Chisholm and Jay O'Brien, at the Francis Crick Institute; A. Yuen, A. Azhar, K. Lau, C. Schwartz, A. Lee and C. Rider for their logistical support for the human exposure study; and staff at the Centre d'expertise et de services G.nome Qu.bec for their sequencing services and support. Data for this study are based on patient-level information collected by the NHS, as part of the care and support of cancer patients. The data are collated, maintained and quality assured by the National Cancer Registration and Analysis Service, which is part of NHS England (NHSE). We extend our thanks to the skilled Cancer Registration Officers (CROs) within the National Disease Registration Service, who abstracted and registered the English tumour and molecular testing data. For the purpose of Open Access, the author has applied for a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. PEACE Consortium members: A. Flanagan, A. Hackshaw, A. Jayaram, A. M. M. Hasan, A. Toncheva, A. Wingate, A. Bunkum, A. Sharp, A. Tookman, A. Murra, A. Magness, A. Coulton, A. M. Frankell, A. Cluroe, A. Kerr, A. Ortega-Franco, A. Lucanas, A. M. Schmitt, A. Furness, A. Rowan, A. Tutt, A. Green, A. Paterson, A. -L. Cattin, A. Fendler, A. Latifoltojar, A. Huebner, A. Thomas, B. T. Alba, B. Chain, B. Naidu, B. Shum, B. Olisemeke, B. Hampton, B. Hanley, B. Tanchel, G. Langman, C. Naceur-Lombardelli, C. Gerard, C. Caldas, C. Mart.nez-Ruiz, C. Dive, C. Stirling, C. Swanton, C. Ferris, C. Lewis, C. Milner-Watts, C. Spencer, C.-w. Lok, C. Bailey, C. Messiou, C. Wilson, C. Puttick, C. Lee, C. P. Marin, C. Alifrangis, C. Richard, C. T. Hiley, D. Hochhauser, D. Wetterskog, D. A. Moore, D. Deng, D. Marrone, D. Enting, D. Josephs, D. Kelly, D. A. Fennell, D. Biswas, D. Papadatos-Pastos, E. Carlyle, E. Provenzano, E. (L.) Karapanagiotou, E. Pintus, E. L. Lim, E. Beddowes, E. Colliver, E. Nye, F. Gishen, F. Gomes, F. H. Blackhall, F. Byrne, F. Athanasopoulou, G. Attard, G. Stamp, G. Middleton, G. D. Stewart, H. Feng, G. Pulman, G. Leone, H. Bridger, H. Shaw, H. Yan, H. Mudhar, H. Bancroft, H. Pallikonda, I. McNeish, I. Proctor, I. Tomlinson, I. Noorani, I. Lobon, J. Bridgewater, J. L. Reading, J. Black, J. Brenton, J. Larkin, J. Spicer, J. K. Rane, J. Biano, J. Nicod, J. Webb, J. L. Quesne, J. A. Shaw, J. Korteweg, K.-K. Shiu, K. Pearce, K. Young, K. S. S. Enfield, K. Kelly, K. Peat, K. Thol, K. G. Blyth, K. Litchfield, K. Allinson, K. Edmonds, K. Chan, K. Dijkstra, K. Grigoriadis, L. Farrelly, L. Grostate, L. Terry, L. Spain, L. Au, L. Pickering, L. Holt, L. Del Rosario, M. Jamal-Hanjani, M. Linch, M. Sivakumar, M. MacKenzie, M. Al Bakir, M. Collard, M. Forster, M. Falzon, M. Mangwende, M. Carter, M. G. Krebs, M. Emmerich, M.

Europe PMC Funders Author Manuscripts

Akay, M. Jimenez-Linan, M. Dietzen, M. Angelova, M. Mitchison, M. O'Flaherty, N. Kanu, N. Magno, N. Yousaf, N. Wright, N. McGranahan, N. Hunter, O. Vainauskas, O. Curtis, O. Lucas, O. Pich, O. Al-Sawaf, P. Prymas, P. Roxburgh, P. Campbell, P. Oliveira, P. Cockcroft, P. Colloby, P. Ellery, P. Hill, P. Parker, P. Van Loo, P. Pawlik, P. Stone, S. Veeriah, R. Vendramin, R. Leslie, R. Fitzgerald, R. Zaidi, R. E. Hynds, R. Salgado, R. Wilson, R. Sinclair, R. Stewart, R. Mitu, S. Beck, S. K. Bola, S. Brandner, S. M. Janes, S. Lise, S. A. Quezada, S. Kuong A. Ung, S. Kadiri, S. Holden, S. Turajlic, S. Gamble, S. Jogai, S. Popat, S. Aitken, S. Benafif, J. Dransfield, S. Howlett, S. Shepherd, S. Ghosh, S. Irshad, S. Phillips-Boyd, S. Baijal, S. Zaccaria, S. Fraser, S. M. Lee, S. Hessey, Y. Ning S. Wong, S. Ward, S. Hazell, T. Enver, T. Karasaki, T. Fernandes, T. Ahmad, T. Sahwangarrom, T. Marafioti, T. B. K. Watkins, T. Maughan, U. Mahadeva, U. McGovern, V. Barbe, W. Drake, W. Hill, W. K. Liu, Y. Summers, Z. Tippu, Z. Rhodes, A. Stewart, A. Alzetani, A. J. Patel, A. Kirk, A. Kerr, A. J. Procter, A. Clipson, A. Rice, A. Bajaj, A. Devaraj, A. Grapa, A. G. Nicholson, A. Robinson, A. Leek, A. Montero, A. Karamani, A. Chaturvedi, A. Nakas, A. Nair, A. Ahmed, A. Osman, A. Sodha-Ramdeen, B. B. Campbell, C. Pilotti, C. Castignani, C. Veiga, C -A. Collins-Fekete, C. Proli, C. Lacson, C. Abbosh, C. E. Weeden, C. R. Lindsay, C. Dick, D. Kaniu, D. Chuter, D. Lawrence, D. R. Pearce, D. Patrini, D. Karagianni, D. Levi, D. G. Rothwell, E. Boleti, E. Borg, E. Kilgour, E. Smith, E. Hoxha, E. L. Cadieux, E. M. Hoogenboom, E. Lim, E. Fontaine, E. Gronroos, F. Granato, F. Galvez-Cancino, F. Monk, F. Fraioli, F. Gimeno-Valiente, G. A. Wilson, G. Royle, G. Kassiotis, G. Stavrou, G. Mastrokalos, G. Price, G. Matharu, H. Zhang, H. Zhai, H. K. Dhanda, H. Bhayani, H. Cheyne, H. Doran, H. L. Lowe, H. Shackleford, H. Chavan, H. Raubenheimer, H. J. W. L. Aerts, I. G. Matos, J. D. Hodgkinson, J. Goldman, J. R. M. Black, J. W. Holding, J. Wilson, J. F. Lester, J. Herrero, J. Richards, J. M. Lam, J. Riley, J. A. Hartley, J. Demeulemeester, J. Tugwood, J. Kisistok, J. Cave, J. Novasio, J. Choudhary, K. S. Peggs, K. Brown, K. Selvaraju, K. Gilbert, K. M. Kerr, K. Rammohan, K. Ang, K. G. Blyth, K. W. Ng, K. Chen, K. AbdulJabbar, K. Thakkar, K. Bhakhri, L. Ensell, L. Joseph, L. Robinson, L. Primrose, L. Scarlett, L. Ambrose, L. Priest, M. Djearaman, M. Hewish, M. N. Taylor, M. Shah, M. Scarci, M. V. Duran, M. Litovchenko, M. W. Sunderland, M. S. Hill, M. D. Forster, M. Hayward, M. Sokac, M. Carter, M. Thomas, M. J. Shackcloth, M. Leung, M. Escudero, M. Diossy, M. Tani*f*., M. Asif, M. Tufail, M. F. Chowdhry, M. Khalil, M. Scotland, M. Malima, N. Fernandes, N. Navani, N. Totten, N. J. Birkbak, N. Gower, N. Panagiotopoulos, N. Kostoulas, O. Ansorge, O. Chervova, P. Ingram, P. Gorman, P. Ashford, P. Bishop, P. De Sousa, P. Russell, P. Crosbie, P. Hobson, P. Shah, R. Rosenthal, R. Shah, R. Boyles, R. Khiroya, R. K. Stone, R. M. Thakrar, R. Bentham, R. C. M. Stephens, R. Bilancia, R. F. Schwarz, S. Saghafinia, S. Lopez, S. Booth, S. Danson, S. Rudman, S. Dulloo, S. Smith, S. Chee, S. Vanloo, S. Richardson, S. Austin, S. I. Buderi, S. Jordan, S. Begum, S. Rathinam, S. Boeing, S. Bandula, S. Moss, S. K. Bola, T. Denner, T. P. Mourikis, T. L. Kaufmann, T. Cruickshank, T. Clark, V. Spanswick, V. Joshi, W.-T. Lu, X. Pan, Y. Wu, Y. Yuan, Y. Naito, Z. Ramsden, Z. Szallasi, A. Dwornik, G. Langman, H. Shackleford, A. Osman, B. Naidu, K. Bowles, C. Grieco, M. L. Le, P. D. D'Arienzo and E. Turay. E.G was supported by an ERC Advanced grant (PROTEUS, grant agreement no 835297)

Data availability

Duplex-seq data for the PEACE and BDRE cohorts are available at the European Genome-Phenome Archive (EGA) with the identifier EGAS00001006951. Duplex-seq data generated from PEACE study samples during this study are not publicly available and restrictions apply to the availability of these data. Such Duplex-seq data are available through the Cancer Research UK and University College London Cancer Trials Centre (ctc.peace@ucl.ac.uk) for academic, non-commercial research purposes upon reasonable request and subject to review of a project proposal that will be evaluated by a PEACE data access committee, entering into an appropriate data access agreement and subject to any applicable ethical approvals. Duplex-seq data generated from the BDRE study are available through J. DeGregori (James.Degregori@cuanschutz.edu) for academic, non-commercial research purposes upon reasonable request, entering into an appropriate data access agreement and subject to any applicable ethical approvals. The Duplex-seq data for the BDRE and PEACE studies were generated using a larger panel of probes that covered approximately 50 kb of the genome, spanning hotspots frequently mutated in cancers. This full dataset has been provided for the 17 never-smoker individuals from the PEACE study. For all other samples, only data for the EGFR and KRAS regions queried are included in this manuscript. The RNA-seq data for the COPA study are available at the EGA with the identifier EGAS00001006966. De-identified participant data are available upon reasonable request to C.C. (christopher. carlsten@ubc.ca) for academic, non-commercial

Europe PMC Funders Author Manuscripts

research purposes. Data availability is subject to a data access agreement and applicable ethical approvals. Mouse WGS data are available at the European Nucleotide Archive (ENA) with the identifier PRJEB58221 (ERP143287). Mouse RNA-seq data are available at the ENA with the identifier PRJEB59269 (ERP144330). Source data are provided with this paper.

Code availability

Code for analysis of epidemiology, RNA-seq and WGS data and processing of healthy lung tissue are available at Zenodo (https://doi.org/10.5281/zenodo.7705022).

References

1. Berenblum I, Shubik P. A new, quantitative, approach to the study of the stages of chemical carcinogenesis in the mouse's skin. Br J Cancer. 1947; 1: 383–391. [PubMed: 18906316]
2. Cogliano VJ, et al. Preventable exposures associated with human cancers. J Nat Cancer Inst. 2011; 103: 1827–1839. [PubMed: 22158127]
3. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—a different disease. Nat Rev Cancer. 2007; 7: 778–790. [PubMed: 17882278]
4. Midha A, Dearden S, McCormack R. EGFR mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutMapII). Am J Cancer Res. 2015; 5: 2892–2911. [PubMed: 26609494]
5. Carrot-Zhang J, et al. Genetic ancestry contributes to somatic mutations in lung cancers from admixed Latin American populations. Cancer Discov. 2021; 11: 591–598. [PubMed: 33268447]
6. Myers R, et al. High ambient air pollution exposure among never smokers versus ever smokers with lung cancer. J Thorac Oncol. 2021; 16: 1858.
7. WHO Global Air Quality Guidelines. Particulate Matter ($PM_{2.5}$ and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization; 2021.
8. Kucab JE, et al. A compendium of mutational signatures of environmental agents. Cell. 2019; 177: 821–836. e16 [PubMed: 30982602]
9. Riva L, et al. The mutational signature profile of known and suspected human carcinogens in mice. Nat Genet. 2020; 52: 1189–1197. [PubMed: 32989322]
10. Moody S, et al. Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. Nat Genet. 2021; 53: 1553–1563. [PubMed: 34663923]
11. Zhang T, et al. Genomic and evolutionary classification of lung cancer in never smokers. Nat Genet. 2021; 53: 1348–1359. [PubMed: 34493867]
12. Jamal-Hanjani M, et al. Tracking the evolution of non-small-cell lung cancer. N Engl J Med. 2017; 376: 2109–2121. [PubMed: 28445112]
13. Chen Y-J, et al. Proteogenomics of non-smoking lung cancer in East Asia delineates molecular signatures of pathogenesis and progression. Cell. 2020; 182: 226–244. e17 [PubMed: 32649875]
14. Frankell AM, et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature. 2023; doi: 10.1038/s41586-023-05783-5
15. Takahashi H, Ogata H, Nishigaki R, Broide DH, Karin M. Tobacco smoke promotes lung tumorigenesis by triggering IKKμ- and JNK1-dependent inflammation. Cancer Cell. 2010; 17: 89–97. [PubMed: 20129250]
16. Im H-B, et al. South Korea Encyclopaedia Britannica. accessed 9 March 2023 https://www.britannica.com/place/South-Korea
17. The Ethnic Group. Executive Yuan. accessed 9 March 2023 https://www.ey.gov.tw/state/99B2E89521FC31E1/2820610c-e97f-4d33-aa1e-e7b15222e45a
18. Huang Y, et al. Air pollution, genetic factors, and the risk of lung cancer: a prospective study in the UK Biobank. Am J Respir Crit Care Med. 2021; 204: 817–825. [PubMed: 34252012]

19. McDaniel Mims B, Grisham MB. Humanizing the mouse immune system to study splanchnic organ inflammation. J Physiol. 2018; 596: 3915–3927. [PubMed: 29574759]
20. Hogg JC, Van Eeden S. Pulmonary and systemic response to atmospheric pollution. Respirology. 2009; 14: 336–346. [PubMed: 19353772]
21. Sutherland KD, et al. Multiple cells-of-origin of mutant K-Ras-induced mouse lung adenocarcinoma. Proc Natl Acad Sci USA. 2014; 111: 4952–4957. [PubMed: 24586047]
22. Choi J, et al. Inflammatory signals induce AT2 cell-derived damage-associated transient progenitors that mediate alveolar regeneration. Cell Stem Cell. 2020; 27: 366–382. e7 [PubMed: 32750316]
23. Strunz M, et al. Alveolar regeneration through a Krt8+transitional stem cell state that persists in human lung fibrosis. Nat Commun. 2020; 11 3559 [PubMed: 32678092]
24. Ryu MH, et al. Impact of exposure to diesel exhaust on inflammation markers and proteases in former smokers with chronic obstructive pulmonary disease: a randomized, double-blinded, crossover study. Am J Respir Crit Care Med. 2022; 205: 1046–1052. [PubMed: 35202552]
25. Ryu, MH. Effects of Traffic-Related Air Pollution Exposure on Older Adults with and without Chronic Obstructive Pulmonary Disease. PhD thesis, Univ of British Colombia; 2021.
26. Nolan E, et al. Radiation exposure elicits a neutrophil-driven response in healthy lung tissue that enhances metastatic colonization. Nat Cancer. 2022; 3: 173–187. [PubMed: 35221334]
27. Dost AFM, et al. Organoids model transcriptional hallmarks of oncogenic KRAS activation in lung epithelial progenitor cells. Cell Stem Cell. 2020; 27: 663–678. e8 [PubMed: 32891189]
28. Hiraiwa K, van Eeden SF. Contribution of lung macrophages to the inflammatory responses induced by exposure to air pollutants. Mediators Inflamm. 2013; 2013 619523 [PubMed: 24058272]
29. Klughammer B, et al. Examining treatment outcomes with erlotinib in patients with advanced non-small cell lung cancer whose tumors harbor uncommon EGFR mutations. J Thorac Oncol. 2016; 11: 545–555. [PubMed: 26773740]
30. Takano APC, et al. Pleural anthracosis as an indicator of lifetime exposure to urban air pollution: an autopsy-based study in Sao Paulo. Environ Res. 2019; 173: 23–32. [PubMed: 30884435]
31. Mirsadraee M. Anthracosis of the lungs: etiology, clinical manifestations and diagnosis: a review. Tanaffos. 2014; 13: 1–13. [PubMed: 25852756]
32. Kunzke T, et al. Patterns of carbon-bound exogenous compounds in patients with lung cancer and association with disease pathophysiology. Cancer Res. 2021; 81: 5862–5875. [PubMed: 34666994]
33. Deprez M, et al. A single-cell atlas of the human healthy airways. Am J Respir Crit Care Med. 2020; 202: 1636–1645. [PubMed: 32726565]
34. Sikkema L, et al. An integrated cell atlas of the human lung in health and disease. Preprint at bioRxiv. 2022; doi: 10.1101/2022.03.10.483747
35. Tate JG, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019; 47: D941–D947. [PubMed: 30371878]
36. Su F, et al. RAS mutations in cutaneous squamous-cell carcinomas in patients treated with BRAF inhibitors. N Engl J Med. 2012; 366: 207–215. [PubMed: 22256804]
37. Yoshida K, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature. 2020; 578: 266–272. [PubMed: 31996850]
38. Yokoyama A, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019; 565: 312–317. [PubMed: 30602793]
39. IARC Working Group on the – Evaluation of Carcinogenic Risks to Humans. Diesel And Gasoline Engine Exhausts And Some Nitroarenes. Vol. 105. World Health Organization; 2014.
40. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Outdoor Air Pollution. Vol. 109. World Health Organization; 2016.
41. Turner MC, et al. Outdoor air pollution and cancer: an overview of the current evidence and public health recommendations. CA Cancer J Clin. 2020; 70: 460–479.
42. Chung KM, et al. Endocrine-exocrine signaling drives obesity-associated pancreatic ductal adenocarcinoma. Cell. 2020; 181: 832–847. e18 [PubMed: 32304665]

43. Ridker PM, et al. Effect of interleukin-1β inhibition with canakinumab on incident lung cancer in patients with atherosclerosis: exploratory results from a randomised, doubleblind, placebo-controlled trial. Lancet. 2017; 390: 1833–1842. [PubMed: 28855077]
44. Crapo JD, Barry BE, Gehr P, Bachofen M, Weibel ER. Cell number and cell characteristics of the normal human lung. Am Rev Respir Dis. 1982; 126: 332–327. [PubMed: 7103258]
45. Doll R, Hill AB. Smoking and carcinoma of the lung Preliminary report 1950. Bull World Health Organ. 1999; 77: 84–93. [PubMed: 10063665]
46. Kennedy SR, et al. Detecting ultralow-frequency mutations by duplex sequencing. Nat Protoc. 2014; 9: 2586–2606. [PubMed: 25299156]
47. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments.NAR Genom. Bioinform. 2021; 3 lqab019
48. Valentine CC III, et al. Direct quantification of in vivo mutagenesis and carcinogenesisusing duplex sequencing. Proc Natl Acad Sci USA. 2020; 117: 33414–33425. [PubMed: 33318186]
49. Eeftens M, et al. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance,PMw and PMcoarse in 20 European study areas; results of the ESCAPE project. Environ Sci Technol. 2012; 46: 11195–11205. [PubMed: 22963366]
50. Van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chainedequations in R. J Stat Softw. 2011; 45: 1–67.
51. Department for Environment Food and Rural Affairs. Modelled Background Pollution Data. 2021. https://uk-air.defra.gov.uk/data/pcm-data#population_weighted_annual_mean_pm25_data
52. British Geological Survey. Radon Data. Indicative Atlas of Radon. 2023. https://www.bgs.ac.uk/datasets/radon-data-indicative-atlas-of-radon/
53. ONS Postcode Directory (Latest) Centroids (Office for National Statistics; 2021). accessed 9 March 2023 https://geoportal.statistics.gov.uk/datasets/ons-postcode-directory-november-2022/about
54. Air Korea. 2021. accessed 9 March 2023 https://www.airkorea.or.kr/web
55. Cancer Registry Statistical Data. National Cancer Center; 2021. https://www.ncc.re.kr/main.ncc?uri=english/sub04_Statistics [accessed 13 March 2023]
56. Taiwan Air Quality Monitoring Network. Environmental Protection Administration. 2022. accessed 23 March 2023 https://airtw.epa.gov.tw/ENG/Default.aspx
57. Politi K, et al. Lung adenocarcinomas induced in mice by mutant EGF receptors found inhuman lung cancers respond to a tyrosine kinase inhibitor or to down-regulation of the receptors. Genes Dev. 2006; 20: 1496–1510. [PubMed: 16705038]
58. Jackson EL, et al. Analysis of lung tumor initiation and progression using conditionalexpression of oncogenic K-ras. Genes Dev. 2001; 15: 3243–3248. [PubMed: 11751630]
59. Schantz MM, et al. Development of two fine particulate matter standard referencematerials (<4 μm and <10 μm) for the determination of organic and inorganic constituents.Anal. Bioanal Chem. 2016; 408: 4257–4266.
60. Chan YL, et al. Pulmonary inflammation induced by low-dose particulate matter exposurein mice. Am J Physiol Lung Cell Mol Physiol. 2019; 317: L424–L430. [PubMed: 31364371]
61. Major J, et al. Type I and III interferons disrupt lung epithelial repair during recovery fromviral infection. Science. 2020; 369: 712–717. [PubMed: 32527928]
62. Pedregosa F, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011; 12: 2825–2830.
63. Desai TJ, Brownfield DG, Krasnow MA. Alveolar progenitor and stem cells in lungdevelopment, renewal and cancer. Nature. 2014; 507: 190–194. [PubMed: 24499815]
64. Bankhead P, et al. QuPath: open source software for digital pathology image analysis. Sci Rep. 2017; 7 16878 [PubMed: 29203879]
65. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs:delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016; 17: 31. [PubMed: 26899170]
66. Ewels PA, et al. The nf-core framework for community-curated bioinformatics pipelines.Nat. Biotechnol. 2020; 38: 276–278.

67. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29: 15–21. [PubMed: 23104886]

68. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with orwithout a reference genome. BMC Bioinformatics. 2011; 12: 323. [PubMed: 21816040]

69. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersionfor RNA-seq data with DESeq2. Genome Biol. 2014; 15: 550. [PubMed: 25516281]

70. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach forinterpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102: 15545–15550. [PubMed: 16199517]

71. Young MD, et al. Single cell derived mRNA signals across human kidney tumors. Nat Commun. 2021; 12 3896 [PubMed: 34162837]

72. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fastand bias-aware quantification of transcript expression. Nat Methods. 2017; 14: 417–419. [PubMed: 28263959]

73. Dekkers JF, et al. Long-term culture, genetic manipulation and xenotransplantation of human normal and breast cancer organoids. Nat P rotoc. 2021; 16: 1936–1965.

74. Muios F, Martnez-Jimnez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. In silico saturation mutagenesis of cancer genes. Nature. 2021; 596: 428–432. [PubMed: 34321661]

Europe PMC Funders Author Manuscripts

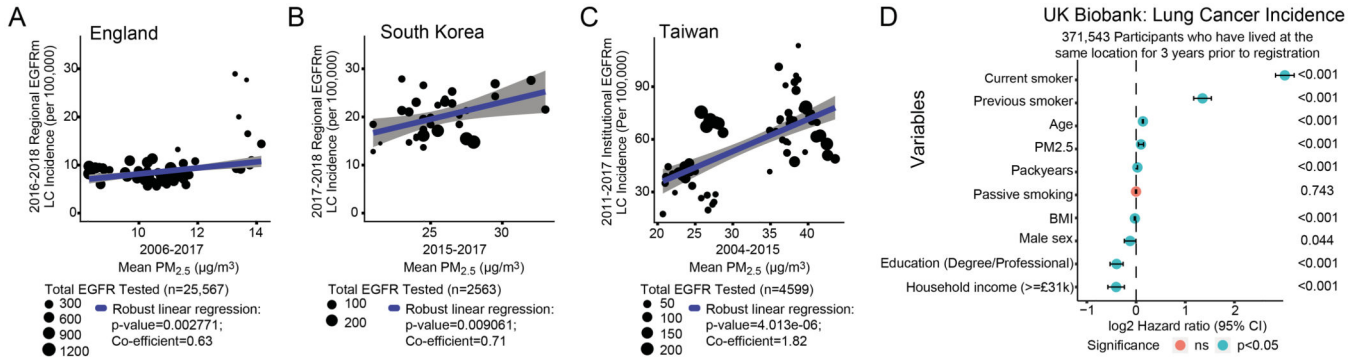


Fig. 1. Exploring the association between cancer and air pollution.

a–c, Scatter plots showing relationships between $PM_{2.5}$ levels and estimated EGFR-driven (EGFR mutant; EGFRm) lung cancer (LC) incidence (per 100,000 population) at the country level in England (a), South Korea (b) and Taiwan (c). Grey shading indicates 95% confidence intervals. d, Forest plot indicating the relationship between lung cancer risk and various co-variates, including residential $PM_{2.5}$ exposure levels (range: 8.17–21.31 μg m^{-3}) in the UK Biobank dataset. Only participants who have lived at the same location for 3 years before registration (n = 371,543) are included. Each co-variate is displayed on a different row. Cox regression P values are indicated on the right. BMI, body–mass index; NS, not significant

Europe PMC Funders Author Manuscripts

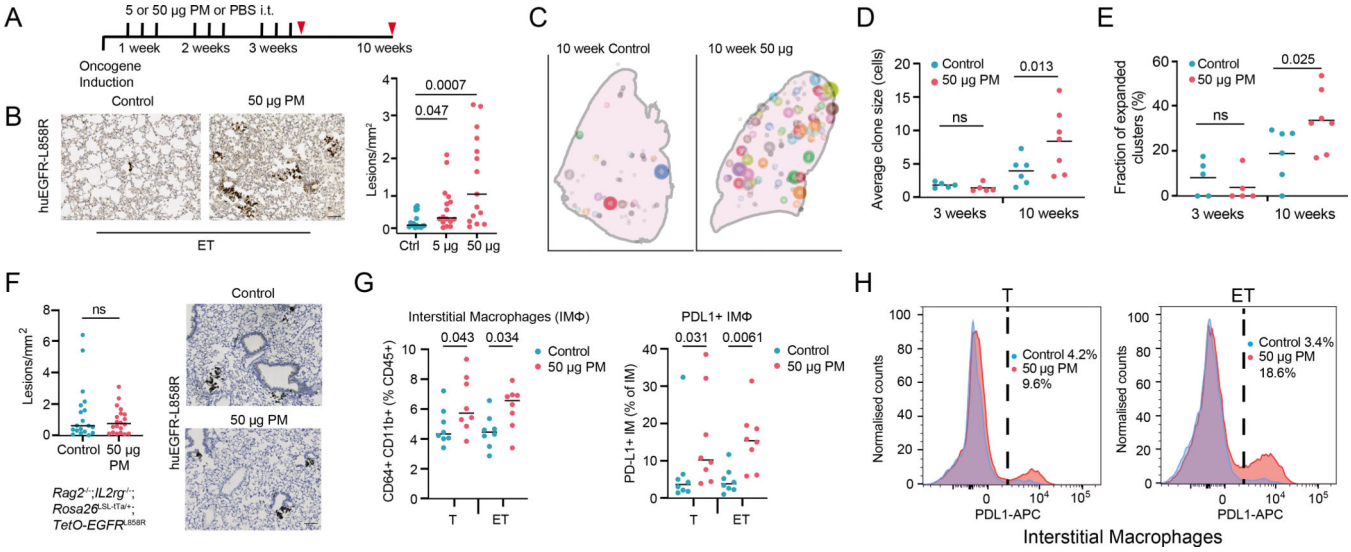


Fig. 2. PM promotes lung tumorigenesis.

a, Schematic of the experiment. Induction of the oncogene was followed by exposure (black lines) through intratracheal (i.t.) administration of PM or PBS (control). Timing of tissue collection is indicated by the red triangles. b, Left, representative immunohistochemistry (IHC) images of human $EGFR^{L858R}$ in ET mice exposed to PBS or PM at 10 weeks. Right, quantification of human $EGFR^{L858R}$-positive neoplasia per mm^2 of lung tissue (n = 16 for the PBS and 5 µg PM groups, n = 15 for the 50 µg PM group). One-way analysis of variance (ANOVA). c, Representative diagram of spatially segmented human $EGFR^{L858R}$-positive clusters in lung lobes, with the size of clusters proportional to $EGFR^{L858R}$ cell number at 10 weeks. d,e, Quantification of average cluster size (d) and fraction of expanded clusters (>5 cells) (e) in mice exposed to PM or PBS over time (n = 5 for 3 week control and 50 µg PM; n = 6, 10 week control; n = 7, 10 week 50 µg PM). One-way ANOVA. f, Left, quantification of lesions in $Rag2^{-/-}$;$Il2rg^{-/-}$;$Rosa26^{LSL\text{-}tTa/+}$;TetO-$EGFR^{L858R}$ mice at 10 weeks after EGFR induction. Right, representative IHC images of $EGFR^{L858R}$ (n = 19 for control, n =20 for 50 µg PM). Mann–Whitney test. g, Proportion of interstitial macrophages (IMs) and PD-L1+IMs within lung tissue in $Rosa26^{LSL\text{-}tdTomato/+}$ mice and ET mice determined by flow cytometry 24 h after final PBS (control) or PM exposure (n = 8 per group). One-way ANOVA. h, Representative histogram showing PD-L1 expression within lung IMs in $Rosa26^{LSL\text{-}tdTomato/+}$ (left) and ET (right) mice exposed to control or PM conditions. Scale bar, 100 .m (b, f). Specific P values are indicated on the charts

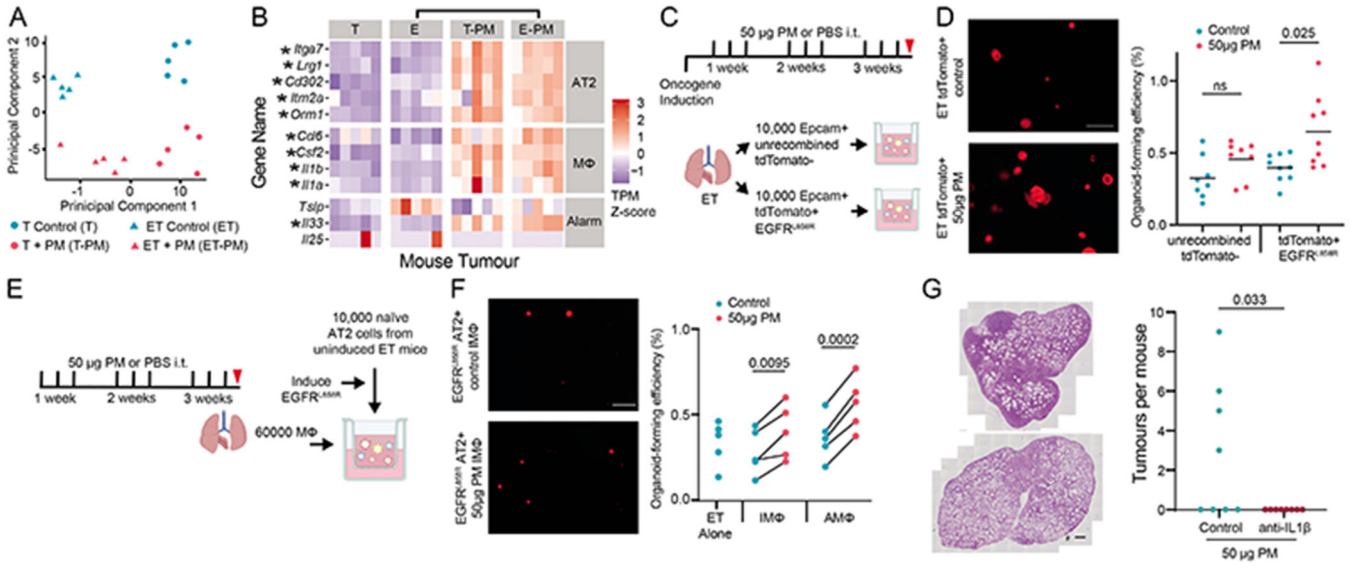


Fig. 3. Increased progenitor-like ability of EGFR mutant cells following PM exposure.

a, PC analysis plot of RNA-seq data from epithelia samples taken from recombined Rosa26$^{LSL\text{-}tdTomato/+}$ mice (T) and ET mice exposed to 50 µg PM or PBS. b, Heatmap of progenitor AT2 cell state (AT2), macrophage recruitment and epithelial alarmin (Alarm) gene expression in all mouse tumour samples. The colour scale in the heatmap represents high (red) to low (blue) transcript per million (TPM) expression z-scores. Asterisks indicate significantly different (FDR < 0.05) gene expression between ET and ET + PM (Methods). c, Schematic of the epithelial organoid assay. Lungs were taken from mice exposed to PM or PBS, followed by isolation and culture of epithelial (positive for epithelial cellular adhesion molecule (EpCAM^{+})) cells. d, Left, representative fluorescent images of tdTomato^{+} organoids at day 14 from ET mice exposed to PBS (control) or PM in vivo. Right, OFE within unrecombined (tdTomato^{-}) or recombined (tdTomato^{+}) EpCAM^{+} lung cells from ET mice exposed to PBS or PM. Two mice were pooled for each biological replicate for sufficient tdTomato^{+} cells. Data represent mean from tdTomato^{-}, n = 8 (16 mice); tdTomato^{+}EGFR^{L858R}, n = 9 (18 mice). One-way ANOVA. e, Schematic of macrophage isolation from mice exposed to PM or PBS and co-cultured with naive (non-PM exposed) EGFR^{L858R} AT2 cells. f, Left, representative fluorescent images of tdTomato^{+}EGFR^{L858R} AT2-cell-derived organoids from ET mice, co-cultured with IMs exposed to PM or PBS. Right, quantification of OFE of EGFR^{L858R} AT2 cells alone and compared with AT2 cells from the same mouse co-cultured with alveolar macrophages (AMs) exposed to PBS or PM (n = 5 mice, data are average of 2 technical replicates per mouse). Paired-t test. g, Left, representative haematoxylin and eosin images of PM-exposed mice treated with IgG control antibody or anti-IL-1β throughout exposure duration. Right, quantification of tumours (n = 8 mice per group). Mann–Whitney test. Scale bar, 500 µm (d,f). The illustrations in c and e were created using BioRender (https://biorender.com)

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

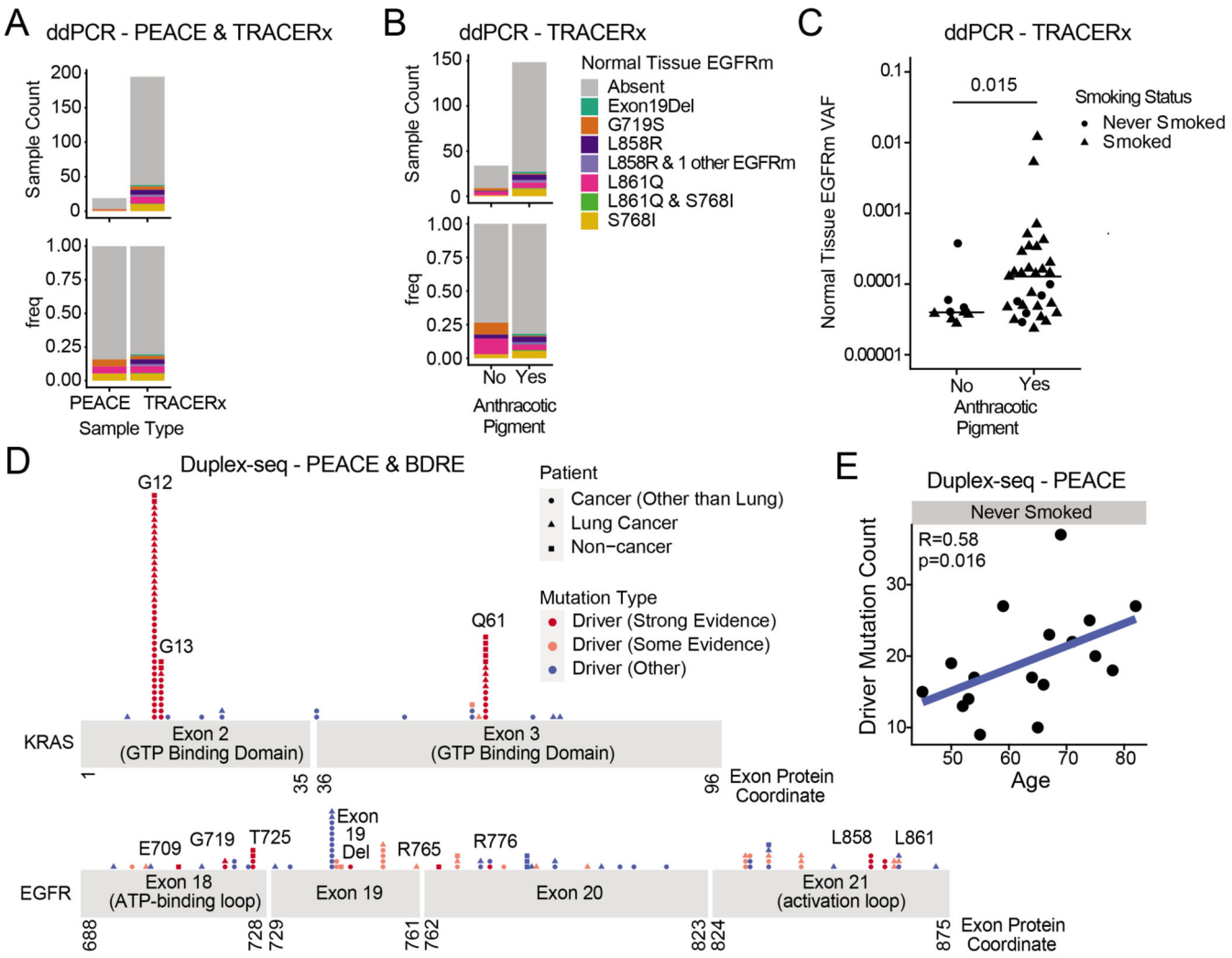


Fig. 4. Mutational landscapes of healthy lung tissue.

a, Counts and proportions of non-cancerous lung samples from PEACE (n = 19) and TRACERx (n = 195) patients that harbour EGFR mutations (EGFRm) identified using ddPCR. The EGFR mutation type is indicated by the colour of the bars (key in b). b, Count and proportion of healthy lung samples from the TRACERx dataset (organized according to anthracotic pigment content: yes (n = 149); no (n = 34)) that harbour EGFR mutations identified by ddPCR. The EGFR mutation type is indicated by the colour of the bars. c, Proportion test Beeswarm plot of ddPCR TRACERx data indicating the VAFs of EGFR mutations. Samples organized according to presence (yes; n = 31) or absence (no; n = 9) of anthracotic pigment. Shapes of dots indicate smoking status. Two-sided t-test. d, Gene models of KRAS (top) and EGFR (bottom), where dots represent mutations identified in the Duplex-seq PEACE and Duplex-seq BDRE cohorts. The position of the dots correspond to the loci of the mutations, whereas the height of the stack indicates the count of the number of mutations at a particular protein coordinate. The shape of the dot indicates the disease diagnosis of the patient, whereas the colour of the dot indicates the mutation type. e, Scatter plot displaying the correlation between age and the number of driver mutations identified

in samples from never-smoker individuals (n = 17) in the Duplex-seq PEACE cohort, for which the panel comprised genomic loci in 31 genes, including EGFR and KRAS. Spearman correlation coefficient and P value are indicated in the plot

Europe PMC Funders Author Manuscripts