Lung Adenocarcinoma Promotion by Air Pollutants

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

*Authors Contributed Equally
**Authors Contributed Equally
^Authors Supervised the Work
#Corresponding Author

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 Department of Thoracic Surgery and Thoracic Oncology Institute, Peking University People's Hospital, Beijing, China.
4 Department of Hematology and Oncology, Chang Gung Memorial Hospital, Chiayi Branch, Chiayi, Taiwan.
5 Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, 06351, Korea.
6 Dept Pathobiology & Population Sciences, The Royal Veterinary College, Hawkshead Lane, N Mymms, Hatfield, Hertfordshire, AL9 7TL
7 Experimental Histopathology STP, The Francis Crick Institute.
8 Oncogene Biology Laboratory, Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK
9 Department of Molecular Cell Biology and Immunology, Amsterdam UMC, Location VUMC, Amsterdam, The Netherlands
10 National Disease Registration Service (NDRS), NHS Digital. London UK
11 Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, Colorado.
12 Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine.
13 Veterans Affairs Eastern Colorado Healthcare System, Aurora, Colorado.

14 Department of Pathology, University of Colorado Anschutz Medical Campus, Aurora, Colorado.
15 Department of Biomedical Sciences, University of Cagliari, Italy
16 Graduate Institute of Clinical Medical Sciences, Chang-Gung University, Taoyuan, Taiwan.
17 Health Information and Epidemiology Laboratory, Chang-Gung Memorial Hospital, Chiayi 613016, Taiwan
18 Centre for Cancer, Society & Public Health, Comprehensive Cancer Centre, School of Cancer and Pharmaceutical Sciences, King’s College London
19 SAGA Diagnostics AB, Lund, Sweden
20 Division of Oncology, Department of Clinical Sciences, Lund University, Lund, Sweden
21 Tumour–Host Interaction Laboratory, The Francis Crick Institute, London, UK
22 Department of Medicine, Division of Respiratory Medicine, Chan-Yeung Centre for Occupational and Environmental Respiratory Disease, Vancouver Coastal Health Research Institute, UBC, Vancouver, BC, Canada.
23 Cancer Research UK & University College London Cancer Trials Centre, London, UK.
24 Department of Medical Oncology, University College London Hospitals, London
25 Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK
26 Agency for Science, Technology and Research, Singapore
27 University College London Medical School, London, UK
28 David H. Koch Institute for Integrative Cancer Research, Cambridge, MA, USA
29 Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
30 BC Cancer Research Institute, University of British Columbia, Vancouver, BC Canada
31 Comprehensive Cancer Centre, King's College London, London, UK.
32 Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK.
33 Ontario Institute for Cancer Research, Toronto, ON, Canada.
34 National Cancer Institute (retired), Bethesda, Maryland.
35 Cancer Research UK City of London Centre, London, UK

Summary

A complete understanding of how environmental carcinogenic exposures promote cancer formation is lacking. Over 70 years ago, tumour formation was proposed to occur in a two step process: an initiating step which induces mutations in normal tissue, followed by a promoter step which triggers cancer development. Recent evidence has revealed healthy human tissue contains a patchwork of clones harbouring oncogenic mutations.This led us to hypothesise that environmental particulate matter measuring <2.5μm ($PM_{2.5}$), known to be associated with lung cancer risk, might promote lung cancer by acting on pre-existing cells harbouring oncogenic mutations in normal lung tissue. Focusing upon EGFR-driven lung cancer, more common in never- or light-smokers, we observed a significant association between $PM_{2.5}$ levels and the incidence of lung cancer in 371,543 individuals from UK BioBank, resident at the same address for at least 3 years, and for 32,957 EGFR-driven lung cancer cases in Public Health England, Taiwan Chang Gung Memorial Hospital, Korean Samsung Medical Centre and the British Columbia Cancer Research Centre cohort. Functional mouse models revealed that pollution causes an influx of macrophages into lung epithelium and interleukin-1β release, resulting in a progenitor-like cell state within EGFR mutant lung alveolar type II epithelial cells that fuel and promote tumorigenesis, in a process that can be attenuated with anti-interleukin-1β. 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 normal tissue samples, respectively. These findings collectively support a tumour-promoting role for $PM_{2.5}$ acting on mutant clones in normal lung tissue, providing support for public health policy initiatives to address air pollution in urban areas to reduce disease burden.

Main text

Introduction

Barrier organs, such as the lung, are directly impacted by exposure to environmental challenges. Accordingly, over 20 environmental and occupational agents are proven lung carcinogens (IARC, 2015), and exposure to these are of particular relevance in understanding lung cancer in the non-smoking population. Lung cancer in never smokers (LCINS) is the 8th most common cause of cancer death in the UK[1] and has distinct clinical and molecular characteristics compared to lung cancer in smokers[2]. In particular, LCINS frequently harbour EGFR oncogenic mutations which tend to be more frequent in female patients, and in East Asian compared to Western patients[3]. Several plausible factors have been proposed to explain the observed sex and geographical disparities of *EGFR* mutant lung cancer, including germline genetics[4], ethnicity, radon exposure, occupational carcinogen exposure and air pollution[5]. While there is a clear tobacco-associated mutational signature in lung cancer in smokers[6], LCINS and EGFR-mutant lung cancers are characterized as harbouring relatively few mutations[7–11] and LCINS has no mutational signature reflecting a specific environmental exposure[7] suggesting alternative mechanisms of LCINS initiation.

Ambient air pollution in the form of particulate matter (PM) is categorized by size including coarse particles with an aerodynamic-mass median diameter, 2.5-10 μm, (PM_{10}), fine particles <2.5 μm, ($PM_{2.5}$) and ultrafine particles (<0.1 μm, $PM_{0.1}$). PM2.5 is of major importance given that approximately 99% of people live in areas that exceed the WHO guideline $PM_{2.5}$ (<5 μg/m^3). Whilst air pollution levels vary widely between countries, estimates suggest it is the world's fourth leading cause of death, accounting for 6.7 million deaths in 2019, with communities with low socioeconomic status disproportionately exposed to higher concentrations[12]. Air pollution arises from a variety of sources including fossil-fuel combustion and the burning of biomass for cooking, with particulate matter (PM) linked to multiple health effects, including chronic obstructive pulmonary disease and asthma[13].

Traditionally, it is thought that carcinogens cause tumours by directly inducing DNA damage[14]. However, recent data from Balmain and colleagues, suggest that many carcinogens do not cause a detectable DNA mutational signature in tumours following exposure [14,15]. A recent genetic analysis found that mutational signatures do not fully explain the varied geographical incidence of oesophageal cancer[16], and efforts that have profiled LCINS tumour genomes failed to detect a dominant carcinogenic signal of mutations deriving from exogenous sources[7,11,17–19]. Furthermore, in the Sherlock study[7] of lung cancer, exogeneous mutational signatures could only be identified in 3% of 232 LCINS genomes.

An alternative hypothesis for how environmental agents might act is by promoting cancer development from initiated but dormant mutant cells[20]. In the absence of exposure to a promoting agent, DMBA-induced mutations in the skin remain dormant for most of the lifespan of the mouse but rapidly progress following treatment with inflammatory stimulus [21]. In support of the presence of such pre-existing mutant cells in normal tissues in humans, sensitive deep sequencing approaches have revealed mutations in clones within normal tissue from multiple organ sites, a minority of which are known to be driver oncogenic mutations in tumours[22–25].

We hypothesised that air pollution might promote inflammatory changes in the normal tissue microenvironment permitting mutated clones to expand and initiate tumours. To address this, we combined epidemiological evidence with functional pre-clinical mouse cancer models, and clinical cohorts providing access to normal lung tissue, to decipher potential mechanisms of air pollution-induced lung tumour promotion and actionable targets for molecular cancer prevention (Figure 1A).

Results

Frequency of EGFRm lung cancer correlates with $PM_{2.5}$ levels across global datasets

Our recently published analysis of the TRACERx 421 cohort revealed that despite a history of smoking, a minority of LUADs (8%) lacked evidence of smoking-mediated mutagenesis, including 6.4% of LUADs associated with >15 years of smoking[26]. Consistent with this analysis, 7-12% of smokers in the TRACERx 421 cohort do not have a driver SNV that can be attributed to a smoking mutation signature (SBS4/SBS92) (Extended Data Figure 1A). Taken together with work from Balmain and colleagues demonstrating that many environmental carcinogens do not lead to a detectable mutagenic signature[15], the question arises as to how environmental carcinogens might facilitate cancer initiation in the absence of detectable DNA mutagenesis. In order to address this question, we studied EGFR mutant (EGFRm) lung cancer which has a high prevalence in LCINS (in England the probability of having an EGFRm tumour in a LCINS patient is 36-40%) and because of the apparent geographical disparities in its occurrence (Supplementary Table 1-3).

To examine the relationship between air pollution and EGFR mutant 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 (92.06% Caucasian cohort; $PM_{2.5}$ IQR: 9.95-11.2 µg/m3), South Korea (estimated >99% Asian cohort[27]: $PM_{2.5}$ IQR: 24.0-27.0 µg/m3) and Taiwan (estimated >98% Asian cohort[4]; $PM_{2.5}$ IQR: 24.3-38.2 µg/m3). In each country, there was a consistent relationship between $PM_{2.5}$ levels (average concentration per geographical area) and estimated EGFR mutant lung cancer incidence (Figure 1B-D). The relative rates of EGFRm lung cancer incidence (per 100,000 population), per 1ug/m3 increment of $PM_{2.5}$ were: England: 0.63 (p-value=0.0028), Korea: 0.71 (p-value=0.0091), Taiwan: 1.82 (p-value=4.01e-06). In addition, when we restricted the England cohort to adenocarcinoma cases, the relationship remained significant (Extended Data Figure 1B).

We were not able to account for migration of individuals prior to diagnosis of lung cancer. As such, we obtained a female never-smoker, lung cancer (92% adenocarcinoma), cohort from British Columbia, Canada, where PM 2.5 cumulative exposure was individually calculated for each case via a detailed residential history from birth to current address, and input into geographical information System mapping (GIS)[29]. The majority of this cohort (83%) were born outside of Canada and 46.7% were EGFR positive. An analysis of 3-year, high compared with low PM 2.5 cumulative exposure and 20-year high compared with low $PM_{2.5}$ cumulative exposure (Methods) revealed the frequency of EGFRm in lung cancer cases was significantly higher after 3 years of high air pollution exposure. (EGFRm frequency - 3 year: High Pollution: 73%, Low Pollution: 40%, p-value=0.03). Of note, this was not observed after 20 years of high vs low cumulative exposure. 20 year: High Pollution: 50%, Low Pollution: 38%, p-value=0.35) (Extended Data Figure 1C). This suggests that 3 years of exposure may be sufficient for EGFRm lung cancers to arise.

To explore if 3 years of cumulative $PM_{2.5}$ exposure is associated with lung cancer (ICD: C33 and C34) in an independent cohort, we obtained data from 407,509 UK Biobank participants (UKBB), where cancer incidence for 28 cancer types, residential location in the 3 years prior to registration, and residential outdoor $PM_{2.5}$ data for the year 2010 were available.

An analysis including all participants regardless of the consistency of residential location in the 3 years prior to registration demonstrated that $PM_{2.5}$ (calculated at 1 $\mu g/m^3$ increments) was associated with lung cancer incidence (Hazard Ratio = 1.08 (95% confidence interval 1.04-1.12); raw p-value=<0.001, FDR=0.001), consistent with a previous analysis of the

UKBB data from Huang et al[30] (Figure 1E; Supplementary Table S4). By contrast, lung cancer incidence was not associated with outdoor radon levels (HR = 0.96 (0.89 - 1.03); raw p-value=0.262). In addition, interaction tests between ever smoking status and $PM_{2.5}$ exposure suggest that smoking and high $PM_{2.5}$ levels may have a combined effect on lung cancer risk (p-value=0.049). We also noted nominal significance (raw p-value<0.05; FDR>0.05) for lip and oropharyngeal cancer (HR = 1.10 (1.01-1.19); raw p-value=0.023; FDR=0.215) and mesothelioma (HR = 1.11 (1.00-1.24), raw p-value=0.048, FDR=0.339). While the estimated HRs from UKBB analyses are higher than in some population based epidemiological surveys[31], this may reflect, the over-representation of wealthier, never-smoker individuals in UKBB (Methods). Finally, we restricted our analysis to participants resident at the same address in the 3 years prior to registration (n=371,543) and observed that the relationship between lung cancer incidence and $PM_{2.5}$ exposure remained significant (HR = 1.07 (1.03-1.11); p-value=<0.001).

Collectively, these data combined with published evidence demonstrating the relationship between $PM_{2.5}$ and LCINS[29], are consistent with an association between the estimated incidence of EGFR mutant lung cancer and levels of $PM_{2.5}$ exposure, and that at least 3 years of air pollution exposure may be sufficient for this association to manifest.

Air pollution promotes EGFR mutant lung cancer progression in mouse models

We next used genetically engineered mouse models of lung adenocarcinoma to functionally examine if PM exposure promotes lung tumour development. We induced expression of oncogenic human *EGFR*^{L858R} mutations in lung tissue using intratracheal delivery of adenoviral-Cre to mice engineered with *Rosa26*$^{\text{LSL-tTa/LSL-tdTomato}}$*; TetO-EGFR*$^{\text{L858R}}$ alleles (ET mice). Following this, we delivered PBS control or fine PM, collected from an urban environment with certified organic and inorganic components[32] at physiologically relevant doses[32]. Mice were given intratracheal administration of PM or PBS control three times per week for three weeks after the induction of *EGFR*$^{\mathrm{L858R}}$ and tumour burden was assessed at 10 weeks post *EGFR^{L858R}* induction (Figure 2A). In this model, rare, sporadic lung epithelial cells express oncogenic *EGFR* and expand to form pre-invasive neoplasia by 10 weeks (Figure 2A,B). Analysis at 10 weeks of ET mice exposed to PM revealed a significant, dose-dependent increase in the number of EGFR mutant cells that had undergone clonal expansions to form neoplasia (control vs 5 µg p-value=0.047; control vs 50 µg p-value=0.0007; Figure 2B). We further validated that PM is influencing epithelial cell tumour formation by targeting *EGFR*$^{\mathrm{L858R}}$ specifically to alveolar type II (AT2) cells using adenoviral SPC-Cre and exposing mice to 50 µg of PM. Exposure to PM was sufficient to significantly increase the number of AT2 derived neoplasia (Extended Data Figure 2A). Exposure to PM before the induction of EGFR mutation in ET mice using adenoviral CMV-Cre also resulted in an increase in the number of early neoplasia (p-value=0.0241; Extended Data Figure 2B), suggesting that PM exposure before or after oncogene induction is sufficient to promote early EGFR mutant driven carcinogenesis.

We observed that PM also increased the number of adenocarcinomas in the more aggressive *CCSP-rtTa; TetO-EGFR*$^{\mathrm{L858R}}$ model of doxycycline-inducible lung adenocarcinoma (p-value=0.032; Extended Data Figure 2C), as well as the number of hyperplasia in an adenoviral-Cre Kras model of lung cancer (*Rosa26*$^{\mathrm{LSL\text{-}tdTomato/+}}$;*Kras*$^{\mathrm{LSL\text{-}G12D/+}}$ (KT); 5 µg p-value=0.048; 50 µg p-value=0.0087;Extended Data Figure 2D). Together, these data suggest that PM can promote tumour progression in both oncogenic Kras and EGFR-driven models of lung adenocarcinoma.

Next we explored the mechanisms by which PM might promote EGFR mutant lung tumourigenesis. Spatial analysis of clonal dynamics throughout early tumourigenesis in ET mice exposed to PM after induction of EGFR$^{\mathrm{L858R}}$ (Figure 2C, see Methods), indicated that EGFR mutant cell expansion is not observed during PM exposure but manifests in the period after PM cessation (p-value=0.0131, Figure 2D); the fraction of EGFR^{L858R} cells that grew into clusters and the number of cells within these clusters were both significantly increased in PM exposed ET mice at 10 weeks but not at 3 weeks (p-value=0.253, Figure 2E). These data suggest PM acts in two ways to promote early tumourigenesis; by increasing the number of EGFR mutant cells with the potential to form a tumour and by elevating the proliferation rate of EGFR mutant cells within these early tumours.

To test if PM promotes tumourigenesis through DNA mutagenesis within epithelial tumour cells, we performed whole genome sequencing (WGS) on tumours from ET mice exposed to 50ug of air pollution (n=5), and PBS controls (n=5). We did not observe a significant increase in the number of mutations in tumours from pollution exposed mice (p-value=0.304), nor enrichment in established single base substitution (SBS) signatures (p-value=0.989), suggesting that short term exposure to PM does not enhance mutagenesis (Extended Data Figure 3). The majority of the mutations in tumours from pollution exposed and control mice were attributable to the ageing signature (Extended Data Figure 3). We next examined if the immune system was required for PM-enhanced EGFR mutant tumourigenesis. We crossed *Rosa26*$^{LSL\text{-}tTa}$ *;TetO-EGFR*^{L858R} mice with *Rag2*$^{-/-}$; *Il2rg*$^{-/-}$ mice which lack T, B, NK cells and have an altered myeloid compartment[33] to generate immune-deficient EGFR mutant mice upon intratracheal delivery of adenoviral Cre (*Rag2*$^{-/-}$*; Il2rg*$^{-/-}$ *;Rosa26*$^{LSL\text{-}tTa/+}$*; TetO-EGFR*^{L858R}). Unlike in the ET mice (Figure 2B), 3 weeks of exposure to PM did not result in a significant increase in neoplasia following *EGFR*^{L858R} induction compared to PBS control mice, suggesting a competent immune system is required for PM-enhanced EGFR mutant lung tumourigenesis (p-value=0.879; Figure 2F).

The inhalation of toxic particles induces a local response in the lung which is mediated by macrophages as well as lung epithelial cells[34,35]. We profiled the acute myeloid response to PM in immune competent lungs harbouring EGFR mutant cells (ET mice) or control (T mice, *Rosa26*$^{LSL\text{-}tdTomato/+}$) 24 hours after the final PM exposure. We observed an increase in the proportion of interstitial macrophages (IMs)(T p-value=0.0427, ET p-value=0.0335 Figure 2G,) and the expression of PD-L1 upon these cells in both T and ET mice following 50μg PM exposure (T p-value=0.0309, ET p-value=0.0061; Figure 2H). Following PM exposure, there was no difference in the proportion of alveolar macrophages (AMs) in the lung (Extended Data figure 4A). There was a significant increase in neutrophils in T mice only and dendritic cells were only elevated in ET mice (Extended Data Figure 4A). Immunofluorescence staining of ET lungs with the pan-macrophage marker CD68 revealed increased density of CD68+ macrophages with PM exposure both acutely and 7 weeks post exposure (3 weeks p-value=<0.0001; 10 weeks p-value=0.0217; Figure 2I). These data suggest transient treatment with PM leads to a sustained increase of PM-associated macrophages throughout early tumorigenesis. We also observed this sustained increase in macrophages in both the doxycycline inducible EGFR^{L858R} model and the KT model at 10 weeks post induction (Extended Data Figure 4B,C) and confirmed these were CD11b+; CD68+ interstitial macrophages (Extended Data Figure 4D). These data support the hypothesis that transient PM exposure is associated with enhanced and sustained lung macrophage infiltration, beyond the time of PM exposure.

Elevated progenitor-like ability of EGFR mutant AT2 cells upon PM exposure

Next, to understand how PM affects lung epithelium, we carried out RNA-seq of flow purified lung epithelia following exposure to four conditions; reporter T mice exposed to PM (T-PM) or PBS control (T), and ET mice exposed to PM (ET-PM) or PBS control (ET). Using principal components (PC) analysis we observed that PM induced significant alterations in the epithelial transcriptome from both T and ET mice, with PM accounting for 19% of the variance in differentially expressed genes (genes differentially expressed between T-PM and T display higher PC2 ranks, p-value<0.001) and EGFR mutation accounting for 38% of the variance (genes differentially expressed between ET and T display higher PC1 ranks, p-value<0.001) (Figure 3A), (Supplementary Table S5). Gene set enrichment analysis of ET mice exposed to PM compared to ET control mice revealed that IL6-JAK-STAT, inflammatory response and allograft rejection pathways were uniquely upregulated upon exposure to PM in EGFR-mutant epithelium in comparison to T mice (Figure 3B; Extended Data Figure 5A). In particular, we observed upregulation of genes known to regulate macrophage recruitment (interleukin-1β (IL1β), GM-CSF, CCL6 and NF-Kb) and the epithelial-derived alarmin (IL33) in PM exposed mouse epithelia (Figure 3C). Lung injury models in mice can induce cell state changes within a proportion of AT2 cells, a likely cell of origin of lung adenocarcinoma[36], and expand populations with a progenitor-like phenotype which mediate alveolar regeneration, and can be driven by inflammatory signals such as IL1β [37,38]. Consistent with our data showing that PM can promote tumorigenesis from EGFR^{L858R} mutant AT2 cells, we noted upregulation of genes previously associated with altered, progenitor-like AT2 cell states in PM treated mouse epithelial tissue (Figure 3C). In addition, deconvolution of single cell signals trained on mouse lung scRNA-seq of bleomycin treated mouse lungs[39] identified a significantly increased Krt8+ AT2 progenitor state score only in ET mice exposed to PM (Extended Data Figure 5B) suggesting EGFR^{L858R} mutant AT2 cells are transcriptionally reprogrammed to this progenitor cell state with PM exposure. We compared the mouse RNA-seq data to a human clinical crossover study, in which lung brushings from people who had never smoked were taken after exposure to diesel exhaust and filtered air[40,41]. A number of gene expression changes, significantly up-regulated in mouse lung epithelium were also up-regulated in human lung epithelium (but not reaching significance in this small human cohort) after PM exposure including IL1β, IL1a markers of macrophage recruitment and ORM1 and LRG1 markers of the AT2 cell state (Extended Data Figure 5C,D). The details of each gene in this comparison are detailed in Supplementary Table S5.

These results identify PM induced inflammatory pathways in mice and humans and transcriptional changes associated with lung progenitor cell states[37]. To test if these transcriptional changes are reflected in functional differences in epithelial cell progenitor behaviour following PM exposure, we performed a lung organoid formation assay[42] in which lung epithelial cells from ET mice were isolated and grown as 3D organoids *ex vivo* following *in vivo* exposure to PM (Figure 3D). Non-recombined cells from ET mice exposed to PM did not display a significant increase in organoid formation efficiency (OFE)(p-value=0.0747; Figure 3E). In contrast, recombined, tdTomato+ *EGFR*^{L858R} cells exposed to PM demonstrated a more pronounced and significant increase in OFE (p-value=0.0245; Figure 3E). To validate whether specifically AT2 cells are functionally altered by PM, we flow purified AT2 cells from non-induced ET or T mice exposed to PM, according to published protocols[43], and subsequently performed adenoviral-Cre recombination *in vitro*[44] in order to express *EGFR*^{L858R} and *tdTomato* or just the reporter control, before plating in the organoid assay (Extended Data Figure 5E). We observed significantly elevated OFE only in tdTomato+ *EGFR*^{L858R} AT2 cells from mice exposed to PM *in vivo* (p-value=0.0043; Extended Data Figure 5F, G), consistent with our in vivo data (Extended Data Figure 2A,D) demonstrating that the AT2 cell is a PM-vulnerable lung cancer cell of origin and that reversing the order of oncogene mutation initiation and PM exposure does not appear to impact tumour initiation capacity. Immunofluorescence confirmed the organoids expressed markers of Krt8+ SPC+ AT2 progenitor states (Extended Data Figure 5H). These data suggest that the combination of *in vivo* exposure of AT2 cells 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.

We previously observed an enrichment of interstitial macrophages in lung epithelium following PM exposure (Figure 2I). Consistent with these data, we observed an increase in a macrophage-recruitment geneset in PM exposed mouse epithelium (figure 3C). We hypothesised that lung macrophages, which generate inflammatory mediators when exposed to particulate matter[35], might be central to the tumour promotion step. To assess whether pollution exposed macrophages are sufficient to drive increased OFE of non-PM exposed AT2 cells, PBS-treated AT2 cells expressing *EGFR*^{L858R} *ex vivo* were co-cultured with either *in vivo* exposed PM or PBS macrophages (Figure 3F). Both PM-exposed interstitial and alveolar macrophages significantly increased the OFE of EGFR mutant AT2 cells (paired t-test, IMs p-value=0.0095; AMs p-value=0.0002; Figure 3G) suggesting a key mediator of PM-induced inflammation arises from macrophages.

Previous reports suggest IL1β derived from lung macrophages is required for the formation of Krt8+ AT2 progenitor cells after bleomycin injury[37], and we noted up-regulation of IL1β in the mouse transcriptomic data following PM exposure (figure 3C), hence we reasoned IL1β may be a key molecular mediator of this pollutant-driven cell state change. We confirmed IL1β is induced by PM using RNAscope and is predominantly arising in CD68+ macrophages (Extended Data Figure 5 I,J). Next, we explored whether treatment with IL1β is sufficient to promote expansion of EGFR mutant organoids. AT2 cells were isolated from naive ET mice, followed by oncogene activation *in vitro* and plated in the organoid assay with IL1β added to the media. This resulted in an expansion of organoid size, with organoids maintaining expression of Krt8+ SPC+ AT2 progenitor states (Extended Data Figure 5K). Finally, to test the requirement of IL1β for PM-mediated adenocarcinoma formation we initiated oncogene expression in the doxycycline inducible *CCSP-rtTa; TetO-EGFR*^{L858R} model; then exposed them to 50 µg of PM and administered anti-IL1β or control antibody (8mg/kg/week) during this exposure period (FIgure 3H). We found that at 10 weeks post-induction, *EGFR*^{L858R} mutant mice treated with anti-IL1β during exposure to PM had significantly attenuated lung adenocarcinoma formation (Figure 3I). Collectively, these data suggest PM exposed macrophages are sufficient to drive a progenitor-like state in EGFR mutant AT2 cells, macrophages are a key source of IL1β in response to PM *in vivo* and IL1β signalling is required for the promotion of PM-mediated EGFR mutant lung adenocarcinoma.

EGFR and *KRAS* mutations are common in normal lung tissue

If tumour development does occur via two steps as originally proposed by Berenblum[45], initiation and promotion, this is contingent on histologically normal tissue cells harbouring oncogenic driver mutations[20]. In 15 reported studies involving deep sequencing of human histologically normal tissues from different anatomic sites (n=9380 samples from 380 patients), an oncogenic *EGFR*^{L858R} mutation was only reported in 1 clone from a skin microbiopsy, suggesting these mutations are rare in well-profiled organs such as the skin, oesophagus, bladder and liver. (Supplementary Table S6). Therefore, we sought evidence for *EGFR* driver mutations in normal lung tissue in people with lung cancer, cancers of other organ sites and individuals with no evidence of cancer, using digital droplet PCR (ddPCR) or Duplex-seq (Figure 4A, Extended Data Figure 7, Supplementary Table S7). Specifically, we only considered mutations that were distinct from those present in matched lung tumours for patients with a history of lung cancer.

We selected normal lung tissue from 195 of 1346 prospectively recruited patients in TRACERx (NCT01888601), balancing the cohort for sex (Female n=96; Male n=99), EGFR mutant tumour status (*EGFR* mutant driver n=39; Other *EGFR* mutant n=10; *EGFR* wt n= 146), smoking status (Ever Smoked n=150; Never Smoked n=45), all within the limits of tissue availability (Figure 4A; Supplementary Table S7, Extended Data Figure 7,8A). We used ddPCR to detect the presence of 5 specific oncogenic *EGFR* driver mutations (Exon19del, G719S, L858R, L861Q, S768I (Klughammer et al., 2016)), and to identify possible clonal expansions in normal lung tissue. The achievable limit of detection was 0.004% based on available input DNA (approximately 600ng per assay).

To exclude the presence of clonal or subclonal spatially distinct *EGFR* mutations that may be present in the corresponding matched lung tumour, we performed multi-region deep next generation sequencing of non-small cell lung cancer (NSCLC) from the same patients

(>3000x coverage) of 19 driver genes (including *EGFR*) using the MiSeq platform. We sequenced 751 tumour regions from the 195 tumours (median 3 regions/tumour) with an achievable limit of detection in each tumour region of 0.966% based on a median sequencing depth per region of 3490X and a MiSeq error rate of 0.473%[46].

We filtered out occurrences of the same mutation in both tumour and normal tissue, potentially attributable to contamination from the primary tumour. After filtering, 38/195 (19%) patients harboured activating *EGFR* mutations exclusively in normal lung tissue that were not detectable in tumour tissue (Figure 4A,B). In tumours from these patients with corresponding normal tissue samples harbouring EGFR mutations, we noted clonal driver mutations in other genes: *TP53*, *PIK3CA*, *KRAS*, *ERBB2*, *CDKN2A*, *BRAF*, and *AKT1*. In patient CRUK267, both *EGFR* L858R and *EGFR* L861Q were detected in normal 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 normal lung tissue, even in patients where the same mutations were not selected during NSCLC tumourigenesis.

To examine whether *EGFR* mutations exist in normal lung tissue from people who never develop lung cancer in their lifetime, we profiled 59 normal lung samples (median 3 samples/patient) collected at autopsy from the PEACE (NCT03004755) study - 19 patients who died of other cancers: Melanoma (n=12), Ovarian Cancer (n=1), Renal Cancer (n=3), Sarcoma (n=2), Mesothelioma (n=1) (Figure 4A, Supplementary Table S7, Extended Data Figure 7,8A). An *EGFR* driver mutation was detected in the normal lung of 16% (3/19) patients (Figure 4B). Despite spatially separated multi-region ddPCR analysis of normal tissue in 15 of the 19 patients, in patients where EGFR driver mutations were detected, they were only detected in one region. Based on the frequency of oncogenic EGFR driver mutations identified in PEACE and TRACERx normal lung samples, we estimated the mutation rate of the individual 5 EGFR mutations using Bayesian inference. (Methods) This yielded a rate of 1 in 2,035,000 (95% credible interval: 1 in 805,000 to 1 in 3,040,000).

Combining these rates to obtain an average EGFR mutation rate, allowed us to estimate that an EGFR oncogenic driver mutation would be present in 1 in 554,500 cells (or around 1:600,000 cells).(Methods)

We next addressed whether there was an association of oncogenic *EGFR* mutations within normal tissue and exposure to ambient pollution in this TRACERx cohort. Anthracosis, determined by the presence of anthracotic pigment (Extended Data Figure 8B), can act as a surrogate for exposure to ambient air pollution[47]. We classified anthracosis within the normal tissue lung samples with and without *EGFR* activating mutations (Figure 4C-D). While there was no association between the presence of an *EGFR* driver mutation in normal tissue and anthracosis (Figure 4C, Prop.test p-value=0.39), there was a significant association between anthracosis and elevated variant allele frequencies of *EGFR* driver mutations (Figure 4D, T-test p-value=0.015). Although there is a trend towards enrichment of smokers in the anthracosis position group (Fisher's exact test, p = 0.065), there are several reports[47–49] that suggest that cigarette smoking is not a risk factor for anthracosis. Moreover, in our cohort the degree of anthracosis observed in never smokers and smokers does not differ; suggesting smoking is not associated with anthracosis (p-value=0.43; Extended Data Figure 8C). While there are multiple environmental contributors to anthracosis[47], these data suggest pollutants are not associated with the frequency of activating oncogenic mutations but rather are associated with the expansion of EGFRm clones.

We sought to validate the identification of EGFRm using an independent ultra-deep sequencing platform in additional cohorts of patients (n=81) with and without cancer, addressing whether driver mutations existed at other genomic loci in *EGFR* and *KRAS*. Using Duplex-seq, we analysed an additional 48 normal lung tissue samples from the PEACE study (NCT03004755) (lung cancer n=9; other cancer n=39), and 33 normal lung tissue samples derived from the Biomarkers and Dysplastic Respiratory Epithelium

(BDRE) Study (NCT00900419, Figure 4A, Supplementary Table S7, Extended Data Figure 7). The BDRE Study cohort consisted of patients with suspicious lung nodules who were referred for evaluation by navigational bronchoscopy at the site of the CT detected lesion (involved site). For each patient, a brushing from the contralateral lung, enriched for bronchial epithelial cells (>89%,[50,51]), was taken for research purposes and used as the source of normal tissue for Duplex-seq. From the BDRE Study cohort, we profiled normal samples from 20 patients with confirmed malignancy in the contralateral lung (lung adenocarcinoma n=10 (including 2 never smokers); lung squamous cell carcinoma n=7; other lung cancer n=2; renal cancer n=1) and normal samples from 13 people without a subsequent cancer diagnosis (including 2 never smokers).

Profiling was carried out using Duplex-seq which identifies mutations within the *EGFR* tyrosine kinase domain exons 18, 19, 20, and 21, *KRAS* GTP binding domain exons 2 and 3, and loci from 29 other genes, with a limit of detection of <0.01%. Given the broader spectrum of *EGFR* mutations detected by Duplex-seq across several exons, we only considered mutations featured in the cancer gene census[52], and further filtered mutations by evidence of driver mutation status in the literature (Supplementary Table S8). In 24 of 68 cancer cases where tissue was available, we also performed Duplex-seq or MiSeq on the corresponding tumour tissue to confirm that the mutations present in normal tissue were found exclusively in the normal tissue samples. Based on the Duplex-seq data, 13/81 (16%) samples harboured an *EGFR* driver mutation (E709X, G719X, T725M, Exon 19 del, R765X, R776X, L858R, L861X; Figure 4E, Extended Data Figure 9A), while 43/81 (53%) samples harboured a *KRAS* driver mutation (G12X, G13X, Q61X; Figure 4E, Extended Data Figure 9B,C). BRAF inhibitors, used to treat BRAFm melanomas, are known to promote accelerated growth of clones harbouring RAS mutations[53]. To exclude the possibility of BRAF inhibitor treatment confounding our analysis, we excluded all melanoma patients from analysis and this did not change the percentage of cases harbouring a KRASm 36/68 (53%). Of note, in samples from smokers from the Duplex-seq PEACE cohort, high confidence (var count>=2) KRAS mutation VAFs were significantly higher than EGFR mutation VAFs (Extended Data Figure 9D, p-value=0.012). Moreover, in the 4 cases that harboured high confidence driver mutations in both KRAS and EGFR, VAFs of KRAS mutations were consistently higher than those in EGFR (Extended Data Figure 9D, paired t-test = 0.01513), suggesting that when KRASm clones and EGFRm clones are both present in normal lungs of smokers, KRASm clones may be more highly selected than EGFRm clones.

In summary, Duplex-seq and ddPCR revealed that 54/295 (18%) of normal lung samples harboured an *EGFR* driver mutation, and 43/81 (53%) normal lung samples harboured a *KRAS* driver mutation. We note that a limitation of our profiling strategies is that we have not purified epithelial cells, the initiating cells of lung tumours, and further work would be required to pinpoint which lineages harbour these mutations. From histological analysis, AT2/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[54]. When we compared proportions of samples that harboured *EGFR* or *KRAS* mutations, no significant trends between smoking status or cancer diagnosis was observed (Supplementary Table S12). We addressed whether mutation signals could be deduced that might shed light on the preponderance of EGFRm LCINS in females. Smoking status, sex, anthracosis and age of patients in the ddPCR TRACERx cohort were entered into a multivariable model to determine which best predicted the likelihood of an EGFR mutation present in normal tissue. Female sex demonstrated the strongest association (p-value=0.06; Extended Data Figure 8D). In order to address whether oncogenic mutations accumulate with the natural ageing process we examined driver mutation frequency harboured by all 31 genes (including EGFR and KRAS) in the Duplex-seq panel in the 17 never smoker patients and noted a significant correlation between age and mutation count in the PEACE cohort (Figure 4F) supporting prior work[25,55].

In summary, these data suggest that oncogenic mutations are present in normal tissue at low frequency and increase with age, fulfilling the initiating step of the Berenblum model. PM results in infiltration of macrophages and release of inflammatory mediators into lung epithelium, including IL1β, which augment progenitor activity of AT2 cells only if these cells harbour an activating oncogenic mutation, fulfilling Berenblum's tumour promoter step.

Discussion

70 years ago, Berenblum and Shubik developed the concept of two processes involved in carcinogenesis; tumour initiation and tumour promotion, the latter involving exposure to an inflammatory but non-mutagenic agent. In the absence of a promotion phase, initiated cells remain dormant for most of the lifespan of a mouse[20]. Balmain and colleagues studied squamous cell carcinoma tumour development in the mouse, showing that cancer development is driven by initiated cells, harbouring DMBA-induced oncogenic mutations in histologically normal tissues, with subsequent inflammatory stimulus in the form of TPA drives tumour promotion and overt malignancy[21,56]. A number of risk factors have been identified for LCINS including second-hand smoke, occupational carcinogen exposure, germline genetics[4] and radon exposure[5]. In this study, we explored the paradigm of tumour promotion driven by particulate matter in the development of lung cancer by air pollutants.

Controlled human exposure studies have found acute diesel exhaust exposure can promote airway inflammation[57]. The IARC monographs 105[58] and 109[59] propose that diesel and gasoline engine exhausts, and outdoor pollution induce lung tumours via genotoxicity, induction of oxidative stress and inflammation. In our manuscript, we build on these previous studies and demonstrate that PM can promote the expansion of pre-existing mutant cells via an inflammatory axis with no detectable environmental carcinogenic DNA signature, which may be amenable to targeting to limit the risk of tumour promotion.

Extending previous findings establishing associations between air pollution and lung cancer[30,31], including LCINS[29], we found an association between the frequency of EGFRm lung cancer incidence and rising $PM_{2.5}$ levels in cohorts from England, South Korea, Taiwan and Canada. Moreover, temporal analysis of the Canadian cohort and UKBB suggests that 3 years of $PM_{2.5}$ may be sufficient to increase risk of EGFRm lung cancer relationship.

A limitation of our analysis of the relationship between EGFRm lung cancer and $PM_{2.5}$ is its ecological nature: using aggregate data instead of participant-level data. We also acknowledge that variables associated with EGFRm status could confound our analysis because they may not be fully adjusted for. In particular, in all three within-country cohorts, and in agreement with the literature[2], EGFRm was more frequent amongst females, Asians, and in lung adenocarcinoma cancer patients (Supplementary Tables 1-3). Even so, our lung cancer study cohorts were well balanced for sex, covered geographically and genetically distinct (Caucasian and Asian) populations, and our England analysis remained significant when we restricted the cohort to adenocarcinoma (Extended Data Figure 1B). Moreover, the functional animal models in this study are restricted to EGFRm and KRAS mutant lung adenocarcinoma only.

Consistent with a model in which PM exposure may serve as the promoter for clonal expansions of oncogenic mutations in normal tissues this model, we find driver mutations in *EGFR* and *KRAS* in normal human lung tissue, adding to research identifying mutations within a range of histologically normal tissues[22–25]. These *EGFR* and *KRAS* mutations are found at similar frequencies in normal lung tissue from patients with an established diagnosis of NSCLC and from patients who do not acquire NSCLC in their lifetime.

We observed that PM promotes lung cancer in mouse models and fosters an AT2 progenitor cell state in *EGFR* mutant cells from mice which can be replicated by incubating naïve PBS exposed AT2 cells with PM exposed macrophages. Prior work has shown that the cytokine

IL1β can promote formation and growth of progenitor AT2 cells[37] and we find that blocking IL1β *in vivo* is sufficient to attenuate PM-mediated EGFR mutant lung adenocarcinoma. Although these mouse models will develop adenomas in the absence of PM and likely do not replicate the complex spectrum of mutations found in normal tissue of a healthy adult, they provide controlled environments to allow insight into early tumourigenesis. These results suggest that cells in normal tissue harbouring driver mutations are restrained from tumour progression but PM exposure can promote inflammation and trigger a rare population of 'dormant' cells to adopt a progenitor cell state, expand and initiate tumourigenesis, as seen by the association of anthracosis and elevated variant allele frequency (VAF) of *EGFR* mutations in normal human lung tissue. The rarity of these mutations in normal tissue (we estimate 1:600,000 cells), combined with the scarcity of the AT2 population and the prolonged requirement for PM exposure in humans may begin to explain the relatively low frequency of EGFRm lung cancer at the population level and the resilience of the lung at the single cell level to cancer initiation.

Our results provide additional evidence that a major risk factor for cancer development is not only the inevitable acquisition of driver mutations in normal epithelium but also mechanisms (both intrinsic and extrinsic) that promote nascent mutant cell expansion and progenitor activity. Assuming little can be done to prevent the inexorable acquisition of oncogenic mutations in normal tissues with age, attention must be turned to addressing the mechanistic, DNA mutation-independent, causes of environmental carcinogenesis.

Balmain and colleagues have demonstrated that most environmental carcinogens tested do not induce a DNA mutagenic signature; 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. TRACERx has revealed that 8% of LUADs in smokers have no detectable smoking carcinogenic signature[26]. Further work should investigate the possibility that tobacco exposure might promote lung cancer through non-mutagenic mechanisms. E-cigarettes should also be evaluated for their potential to generate inflammatory responses in the lung necessary for the promotion step in the Berenblum model. Indeed there is an urgent need for carcinogenic assays that effectively reflect the potential for tumour promotion across different tissues and to understand tissue-specific inflammatory mediators of this process.

Such efforts may guide novel screening paradigms in high-risk, under-served populations and "molecularly targeted" cancer prevention approaches to inhibit cancer initiation. It is notable that the antibody Canikumumab, against one such "promoter" target, IL1β, induced in both mouse and human following PM exposure has already been shown to reduce lung cancer incidence in the cardiovascular prevention trial, CANTOS[60].

In conclusion, our data suggest a mechanistic and causative link between pollution and lung cancer, first proposed by Doll and Hill in 1950[61], providing a public health mandate to urgently restrict particulate emissions in urban areas.

Main Text References

1. Bhopal, A., Peake, M. D., Gilligan, D. & Cosford, P. Lung cancer in never-smokers: a hidden disease. *J. R. Soc. Med.* **112**, 269–271 (2019).ision
2. Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers--a different disease. *Nat. Rev. Cancer* **7**, 778–790 (2007).
3. 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.* **5**, 2892–2911 (2015).
4. Carrot-Zhang, J. *et al.* Genetic Ancestry Contributes to Somatic Mutations in Lung Cancers from Admixed Latin American Populations. *Cancer Discov.* **11**, 591–598 (2021).
5. Couraud, S., Zalcman, G., Milleron, B., Morin, F. & Souquet, P.-J. Lung cancer in never smokers--a review. *Eur. J. Cancer Oxf. Engl. 1990* **48**, 1299–1311 (2012).
6. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
7. Zhang, T. *et al.* Genomic and evolutionary classification of lung cancer in never smokers. *Nat. Genet.* **53**, 1348–1359 (2021).
8. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
9. Devarakonda, S. *et al.* Genomic Profiling of Lung Adenocarcinoma in Never-Smokers. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **39**, 3747–3758 (2021).
10. Govindan, R. *et al.* Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. *Cell* **150**, 1121–1134 (2012).
11. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
12. GBD 2019 Tobacco Collaborators. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study

2019. *Lancet Lond. Engl.* **397**, 2337–2360 (2021).

13. Cohen, A. J. *et al.* Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet Lond. Engl.* **389**, 1907–1918 (2017).
14. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16 (2019).
15. Riva, L. *et al.* The mutational signature profile of known and suspected human carcinogens in mice. *Nat. Genet.* **52**, 1189–1197 (2020).
16. Moody, S. *et al.* Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. *Nat. Genet.* **53**, 1553–1563 (2021).
17. Chen, Y.-J. *et al.* Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. *Cell* **182**, 226-244.e17 (2020).
18. Lee, J. J.-K. *et al.* Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma. *Cell* **177**, 1842-1857.e21 (2019).
19. Wang, C. *et al.* Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. *Nat. Commun.* **9**, 2054 (2018).
20. 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* **1**, 383–391 (1947).
21. Balmain, A. The critical roles of somatic mutations and environmental tumor-promoting agents in cancer risk. *Nat. Genet.* **52**, 1139–1143 (2020).
22. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
23. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
24. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
25. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial

epithelium. *Nature* **578**, 266–272 (2020).

26. Frankell, A., Dietzen, M., Al Bakir, M. & Lim, E. The evolution of lung cancer and impact of subclonal selection in TRACERx. *Nature* **In Press**,.
27. South Korea | History, Map, Flag, Capital, Population, President, & Facts | Britannica. https://www.britannica.com/place/South-Korea.
28. 2.16.886.101.20003. 行政院全球資訊網. *2.16.886.101.20003* https://www.ey.gov.tw/state/99B2E89521FC31E1/2820610c-e97f-4d33-aa1e-e7b15222e45a (2011).
29. Myers, R. *et al.* High Ambient Air Pollution Exposure Among Never Smokers Versus Ever Smokers with Lung Cancer. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* S1556-0864(21)02256–5 (2021) doi:10.1016/j.jtho.2021.06.015.
30. 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.* **204**, 817–825 (2021).
31. Turner, M. C. *et al.* Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *CA. Cancer J. Clin.* (2020) doi:10.3322/caac.21632.
32. Schantz, M. M. *et al.* Development of two fine particulate matter standard reference materials (<4 μm and <10 μm) for the determination of organic and inorganic constituents. *Anal. Bioanal. Chem.* **408**, 4257–4266 (2016).
33. McDaniel Mims, B. & Grisham, M. B. Humanizing the mouse immune system to study splanchnic organ inflammation. *J. Physiol.* **596**, 3915–3927 (2018).
34. Hogg, J. C. & Van Eeden, S. Pulmonary and systemic response to atmospheric pollution. *Respirology* **14**, 336–346 (2009).
35. Hiraiwa, K. & van Eeden, S. F. Contribution of Lung Macrophages to the Inflammatory Responses Induced by Exposure to Air Pollutants. *Mediators Inflamm.* **2013**, 619523 (2013).

36. Sutherland, K. D. *et al.* Multiple cells-of-origin of mutant K-Ras–induced mouse lung adenocarcinoma. *Proc. Natl. Acad. Sci.* **111**, 4952–4957 (2014).

37. Choi, J. *et al.* Inflammatory Signals Induce AT2 Cell-Derived Damage-Associated Transient Progenitors that Mediate Alveolar Regeneration. *Cell Stem Cell* **27**, 366-382.e7 (2020).

38. Zacharias, W. J. *et al.* Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor. *Nature* **555**, 251–255 (2018).

39. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nat. Commun.* **11**, 3559 (2020).

40. Ryu, M. H. *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.* **205**, 1046–1052 (2022).

41. Ryu, M. H. Effects of traffic-related air pollution exposure on older adults with and without chronic obstructive pulmonary disease. (University of British Columbia, 2021). doi:10.14288/1.0398486.

42. Nolan, E. *et al.* Radiation exposure elicits a neutrophil-driven response in healthy lung tissue that enhances metastatic colonization. *Nat. Cancer* **3**, 173–187 (2022).

43. Major, J. *et al.* Type I and III interferons disrupt lung epithelial repair during recovery from viral infection. *Science* **369**, 712–717 (2020).

44. Dost, A. F. M. *et al.* Organoids Model Transcriptional Hallmarks of Oncogenic KRAS Activation in Lung Epithelial Progenitor Cells. *Cell Stem Cell* **27**, 663-678.e8 (2020).

45. Berenblum, I. & Shubik, P. The persistence of latent tumour cells induced in the mouse's skin by a single application of 9:10-dimethyl-1:2-benzanthracene. *Br. J. Cancer* **3**, 384–386 (1949).

46. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).

47. Takano, A. P. C. *et al.* Pleural anthracosis as an indicator of lifetime exposure to urban

air pollution: An autopsy-based study in Sao Paulo. *Environ. Res.* **173**, 23–32 (2019).

48. Mirsadraee, M. Anthracosis of the Lungs: Etiology, Clinical Manifestations and Diagnosis: A Review. *Tanaffos* **13**, 1–13 (2014).
49. Kunzke, T. *et al.* Patterns of Carbon-Bound Exogenous Compounds in Patients with Lung Cancer and Association with Disease Pathophysiology. *Cancer Res.* **81**, 5862–5875 (2021).
50. Deprez, M. *et al.* A Single-Cell Atlas of the Human Healthy Airways. *Am. J. Respir. Crit. Care Med.* **202**, 1636–1645 (2020).
51. Sikkema, L. *et al.* An integrated cell atlas of the human lung in health and disease. 2022.03.10.483747 Preprint at https://doi.org/10.1101/2022.03.10.483747 (2022).
52. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
53. Su, F. *et al.* RAS Mutations in Cutaneous Squamous-Cell Carcinomas in Patients Treated with BRAF Inhibitors. *N. Engl. J. Med.* **366**, 207–215 (2012).
54. Jd, C., Be, B., P, G., M, B. & Er, W. Cell number and cell characteristics of the normal human lung. *Am. Rev. Respir. Dis.* **125**, (1982).
55. Kakiuchi, N. *et al.* Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis. *Nature* **577**, 260–265 (2020).
56. Huang, P. Y. & Balmain, A. Modeling cutaneous squamous carcinoma development in the mouse. *Cold Spring Harb. Perspect. Med.* **4**, a013623 (2014).
57. Long, E. & Carlsten, C. Controlled human exposure to diesel exhaust: results illuminate health effects of traffic-related air pollution and inform future directions. *Part. Fibre Toxicol.* **19**, 11 (2022).
58. IARC Working Group on the & Evaluation of Carcinogenic Risks to Humans. *IARC Monograph 105 - DIESEL AND GASOLINE ENGINE EXHAUSTS AND SOME NITROARENES*. (2014).
59. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. *IARC Monograph 109 - Outdoor Air Pollution*. vol. 109 (2016).

60. Ridker, P. M. *et al.* Effect of interleukin-1β inhibition with canakinumab on incident lung cancer in patients with atherosclerosis: exploratory results from a randomised, double-blind, placebo-controlled trial. *The Lancet* **390**, 1833–1842 (2017).

61. Doll, R. & Hill, A. B. Smoking and carcinoma of the lung. Preliminary report. 1950. *Bull. World Health Organ.* **77**, 84–93 (1999).

Figures

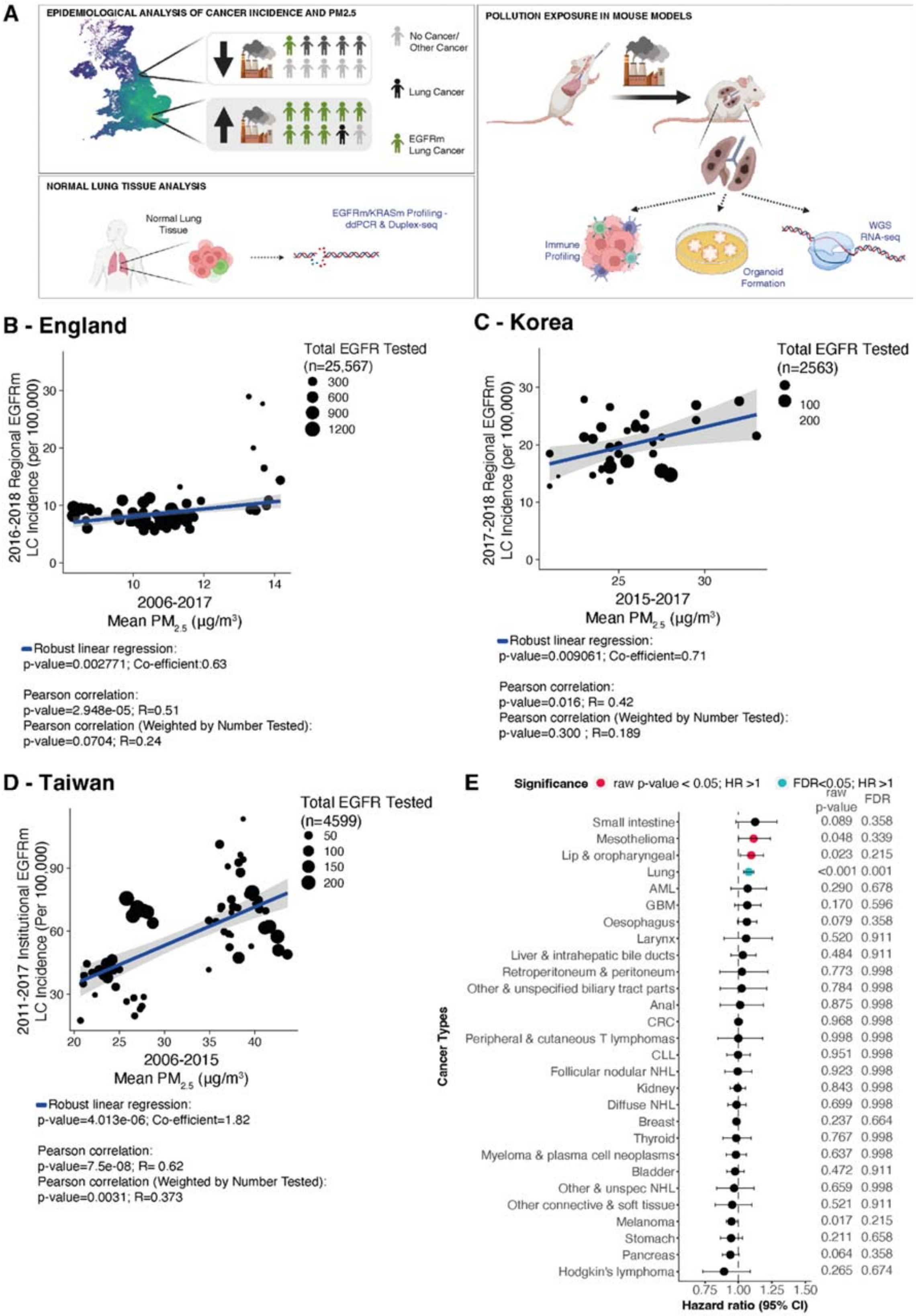

Figure 1: Exploring the association between cancer and air pollution. A) Study design. B-D) Scatter plots showing relationships between $PM_{2.5}$ and estimated *EGFR* mutant lung cancer incidence (per 100,000 population) at the country level in England (B), Korea (C) and Taiwan (D). The blue line

indicates the robust linear regression line. E) Forest plot indicating the relationship between cancer risk and residential $PM_{2.5}$ exposure levels (range: 8.17 - 21.31 μg/m^3) in the UK Biobank dataset (n=407,509). Each cancer type is displayed on a different row. Both raw p-values and FDR values are provided. The color of the dots indicates the level of significance.

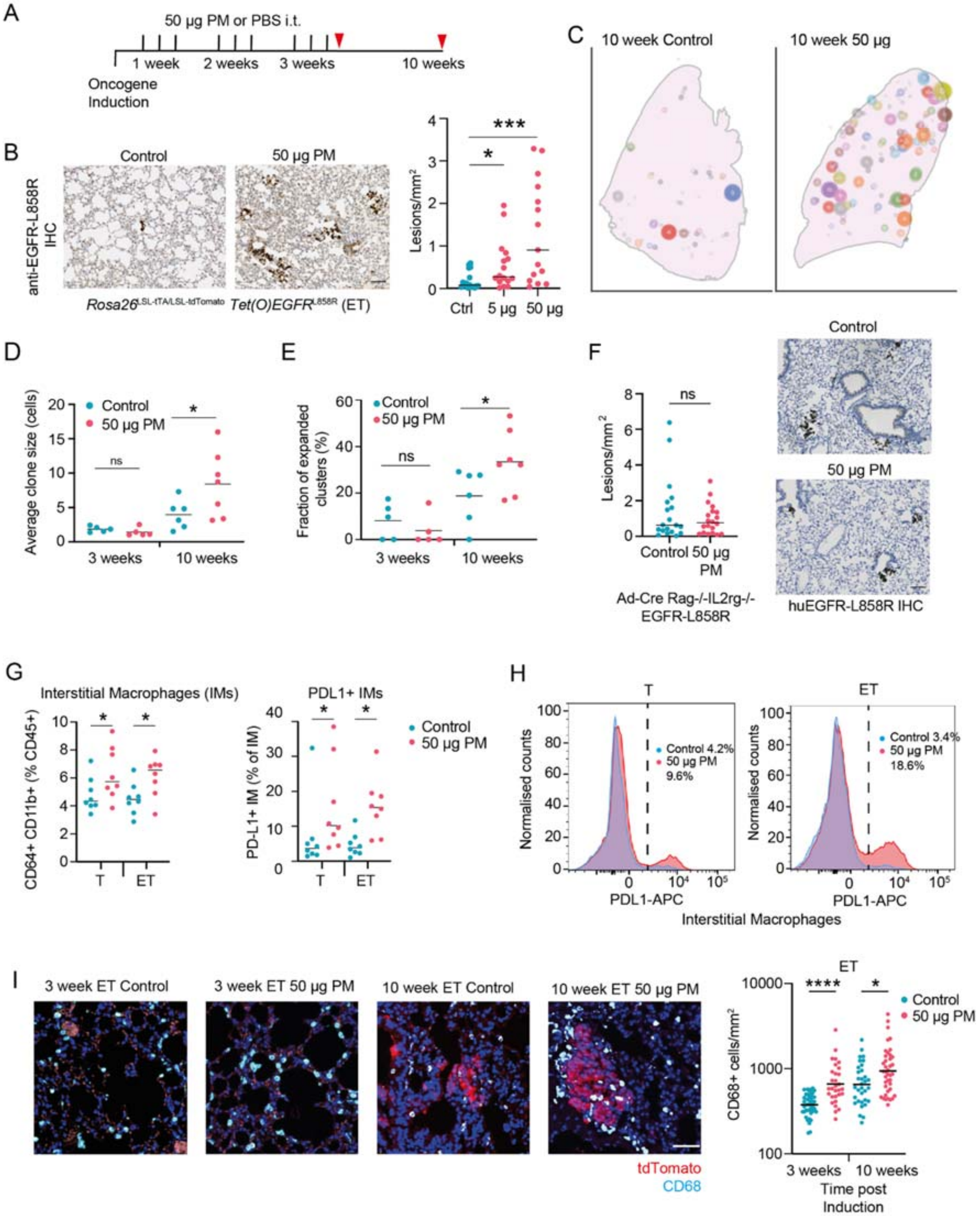

Figure 2: PM promotes lung tumourigenesis. A) Schematic of mouse model of lung cancer indicating induction of oncogene, followed by exposure (black lines) to particulate matter (PM) and tissue collection (red triangles). B) LEFT: Representative immunohistochemistry (IHC) of human EGFR^{L858R} in control and PM exposed ET mice. RIGHT: quantification of huEGFR^{L858R+}

neoplasia/mm^2 of lung tissue (n=16 control & 5 µg group, n=15 for 50 µg group). C) Representative diagram of spatially segmented clusters in control and PM exposed ET lungs at 10 weeks, lung lobe outlined in grey and size of cluster colour is proportional to EGFR^{L858R} cluster size. Quantification of average cluster size (D) and fraction of expanded clusters (>5 cells) (E) in PM and control mice at 3 and 10 weeks. F) LEFT: Quantification of lesions in control and PM exposed Rag-/-; IL2rg-/- EGFR^{L858R} mutant mice at 10 weeks post induction and RIGHT: representative EGFR^{L858R} IHC. G) Proportion of interstitial macrophages (IM's) and PDL1+ IM's within lung tissue determined by flow cytometry in T and ET mice 24 hours after final control (blue) or PM (pink) exposure, (n=8 per group). H) Representative histogram showing PD-L1 expression within lung interstitial macrophages in T(left) and ET (right) mice in control (blue) or PM-exposed (pink) conditions. I) LEFT: Representative immunofluorescent images of CD68+ macrophages (cyan) and tdTomato+ EGFR mutant cells (red) within ET lungs exposed to control or 50 µg PM either 3 weeks (left panel) or 10 weeks (right panel) post oncogene induction. RIGHT: Quantification of CD68+ cells per mm^2 of lung tissue, selecting >30 random fields of view of 500 µm^2 (n= 4 mice per group). Gating strategies for flow cytometry analysis provided in Extended Data Figure 6. Statistical analysis by one-way ANOVA for B, D, E, F, G & I. Scale bars 100 µm (B,F,E), *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

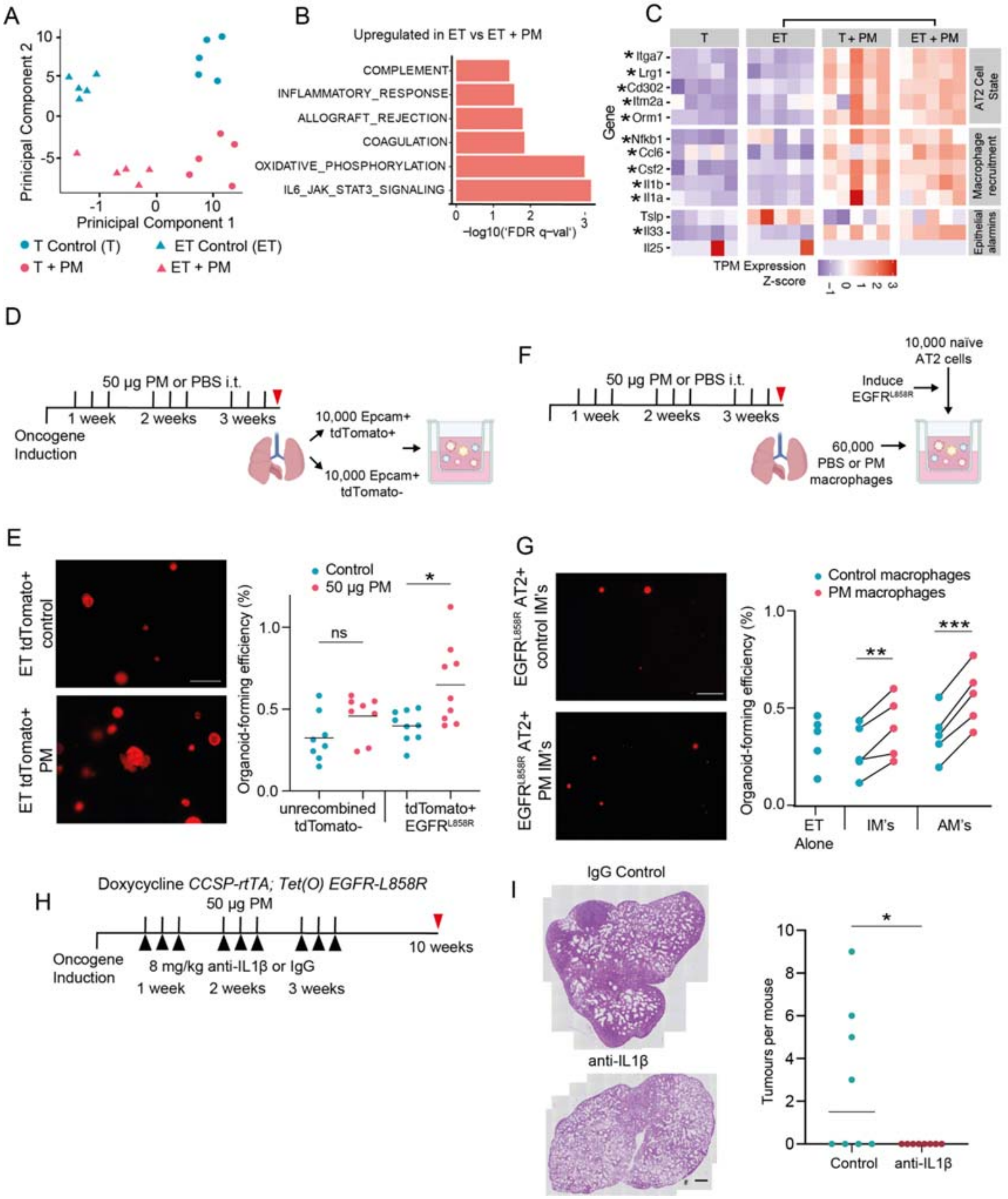

Figure 3: Elevated progenitor-like ability of EGFRm cells upon PM exposure. A) Principal component analysis plot of RNA-seq of epithelia from recombined T and E mice either exposed to PM or control. B) Significantly enriched GSEA pathways upregulated in ET-PM lung epithelial cells compared to ET control mice. C) Heatmap of progenitor AT2 cell state markers, inflammatory, and alarmin gene expression in all samples. The colour scale in the heatmap represents high (red) to low (blue) TPM expression z-scores; asterisks indicate significantly different gene expression between ET and ET + PM (black line). D) Schematic of epithelial organoid assay showing harvesting of lungs from mice exposed to PM or PBS followed by isolation and culture of epithelial (Epcam+) cells. E) LEFT: Representative fluorescent images of tdTomato organoids at day 14 from control ET mice or ET mice exposed to pollution in vivo and RIGHT: organoid forming efficiency (2 mice were pooled for each biological replicate for sufficient tdTomato+ cells: tdTomato- n=8 (16 mice); tdTomato+EGFR n=9 (18 mice)). F) Schematic of isolation of macrophages from mice exposed to PM or PBS and culture with naive (non-PM exposed) EGFR^{L858R} AT2 cells. G) LEFT: Representative fluorescent images of tdTomato AT2 derived organoids co-cultured with PM or PBS exposed macrophages and RIGHT:

quantification of organoid forming efficiency of EGFR mutant AT2 cells alone or with macrophages compared to AT2 cells from the same mouse c-cultured with PM-exposed macrophages. H) Schematic of anti-IL1β treatment treatment (black triangles) during PM exposure (black lines) and harvest (red triangle). I) LEFT: Representative H&E images of PM exposed mice treated with IgG control antibody or anti-IL1B , RIGHT: quantification of tumours (n= 8 mice per group). Statistical analysis by one-way ANOVA for F; paired-t test for G and Mann-Whitney for I. Scale bar 500 μm (F, G & I). *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001

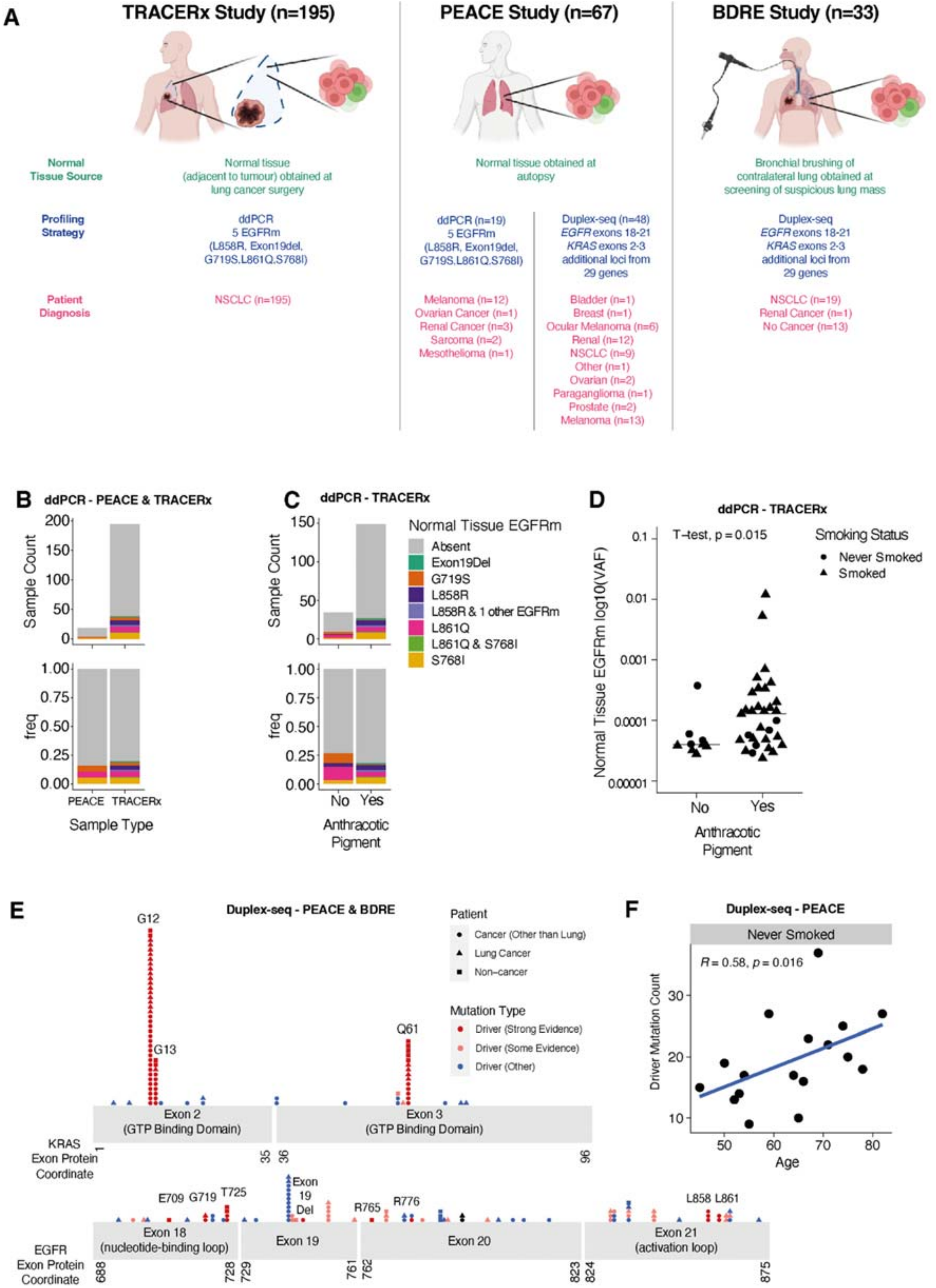

Figure 4: Mutational landscapes of normal lung tissue. A) Schematic indicating normal lung tissue cohorts analysed by ddPCR and Duplex-seq. B) The counts and proportions of PEACE and TRACERx normal lung samples that harbour *EGFR* mutations identified using ddPCR. The *EGFR* mutation type is indicated by the colour of the bars. C) The count and proportion of TRACERx normal lung samples (organised according to anthracotic pigment content) that harbour *EGFR* mutations identified by ddPCR. The *EGFR* mutation type is indicated by the colour of the bars. D) Beeswarm

plot indicating the variant allele frequencies of *EGFR* mutations, samples organized according to presence (yes) or absence (no) of anthracotic pigment. Shapes of dots indicate smoking status. E) 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 while 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 diagnosis of the patient, while the color of the dot indicates the mutation type. G) Scatter plot displaying correlation between age and number of driver mutations identified in the never smoker samples (n=17) in the Duplex-seq PEACE cohort, where the panel comprised genomic loci in 31 genes, including EGFR and KRAS.

Extended Data Figure Legends

Extended Data Figure 1: A) TX421 Tumours from Smokers. Barplots indicating proportion of SNVs in each tumour attributed to each SBS mutational signatures. The barplots (Top: LUAD, Bottom: 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 aging). 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 the probabilities are aggregated. Asterisks represent patients where the smoking-related aggregated probabilities are below 0.5.B) Correlation between PM2.5 levels and EGFRm Adenocarcinoma lung cancer incidence in England. C-D) The Canadian Lung Cancer Cohort. C) Distribution of 3 year and 20 year cumulative PM2.5 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/m3) and 5th quintiles (7.27ug/m3) of the distribution. The full distribution is displayed in the top plot, while the bottom plot displays a narrower range of 4-10 ug/m3 (for clarity). D) Counts and frequencies of EGFRm in the Canadian Cohort, where 3 year and 20 year cumulative PM2.5 exposure levels were available. Patients are grouped into high, intermediate and low groups based on thresholds established as described in (C). These groups are defined based on 3 year cumulative PM2.5 exposure data (left) and based on 20 year cumulative PM2.5 exposure data (right). The bar plots display the counts and frequency of EGFRm amongst patients within each group. The frequency of EGFRm is significantly higher in the high pollution exposure group when compared to the low pollution exposure group only based on 3 year cumulative PM2.5 exposure data but not based on 20 year cumulative PM2.5 exposure data.

Extended Data Figure 2: A) Schematic of PM exposure and representative IHC of ET mice induced with AT2-specific SPC-Cre exposed to PM or PBS control and quantification of neoplastic lesions (n=14 PBS, n=11 PM) B) Schematic of PM exposure followed by induction of EGFR and quantification of precancerous lesions/mm^2 of lung tissue (n=9 PBS; n=8 5µg; n=11 50µg; p=0.0241). 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*^{L858R} mice; quantification of number of adenocarcinomas per mouse below (n = 9 per group). D) Schematic of PM exposure and representative IHC for red fluorescent protein (RFP, marks tdTomato+ cells) in *Rosa26*$^{LSL\text{-}tdTomato/+}$;*Kras*$^{LSL\text{-}G12D/+}$ 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). Scale bar 50 µm (C main, H), 20 µm (C insert), 100 µm A & D.

Extended Data Figure 3: WGS analysis of tumours from mice exposed to air pollution (n=5) and those exposed to PBS controls (n=5). A) Displays mutational profiles for each tumour sample according to the mutation trinucleotide context. 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 (Pollution) and tumours from PBS exposed mice (PBS), All mutations are summarised in one plot on the left, and are then further divided based on the base change of the mutation. 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 aging related signatures (SBS40, SBS5, SBS1).

Extended Data Figure 4: A) Immune cell frequencies in the lungs estimated by FACS 24 hours post-exposure from tdTomato (T) and EGFR mutant (ET) mice after 50µg (red) or control (blue) (n=8 mice per group). Data are presented as the frequency among live immune cells. Representative immunofluorescent images of CD68+ macrophages (cyan) and tdTomato+ Kras mutant cells (red) within KT lungs (B) or *CCSP-rtTA; TetO EGFR^{L858R}* lungs (C) exposed to control or 50 µg PM 10 weeks post oncogene induction and quantification of CD68+ cells per mm^2 of lung tissue, selecting >20 random fields of view of 500 µm^2 (n= 3 mice per group). Scale bar 50 µm B & C, 150 µm D.

Extended Data Figure 5: A) Significantly enriched GSEA pathways upregulated in T-PM lung epithelial cells compared to T control mice. B) Progenitor-like AT2 score based on deconvolution of bulk RNA-seq of T and ET mice exposed to 50 µg PM or PBS. C) Schematic displaying experimental set-up of clinical exposure study in never-smoker volunteers initially reported in [38], crossover design with (i) and (ii) in random order separated by 4-week washout. D) Fold change (FC) of significantly upregulated genes (identified in mouse) compared to the fold change of genes changed in the clinical exposure study. With common directionality across species indicated (negative: grey background; positive: red background). E) Schematic of AT2 culture from E or ET mice exposed to 50 µg PM or PBS. F) Representative fluorescent images of tdTomato organoids at day 14 from E or ET mice exposed to pollution in vivo. G) Quantification of AT2 organoid forming efficiency. n=4 mice per group T and n=5 mice per group ET.H) Fluorescent imaging of Keratin8+ (magenta), SPC+ (blue) AT2 organoids. I) Quantification of IL1B RNAscope and representative IHC. J) Quantification of IL1β positive CD68+ cells at 3 weeks post induction in ET mice following exposure to PM and representative image of IL1B RNAscope (green) in CD68 positive (red) macrophages, arrows indicate positive macrophages. n=3 mice per group and error bar is s.d K) Representative fluorescent images of EGFR-L858R+ AT2 organoids from ET mice treated with control or IL1β in vitro. tdTomato (yellow) organoids stained with SPC (blue) and Keratin 8 (magenta). Scale bar 100µm. Quantification of organoid size with each dot representing an organoid at day 14 of control (blue) or IL1β treated (orange). n=3 mice per group. Scale bar 100 µm F; 20 µm H, I; 50 µm J, K.

Extended Data Figure 6 Gating Strategy: A, B) Example of FACS gating strategy to determine frequency of (A) alveolar macrophages, interstitial macrophages, neutrophils, dendritic cells and (B) epithelial cells both tdTomato positive and negative. All samples were gated to exclude debris and doublets, followed by live cell descrimination. C) Representative picture from a tdTomato treated with PBS via i.t for 3 weeks using sort strategy for AT2 cells defined in Major et al., 2020 and macrophages defined in Choi et al., 2020.

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

Extended Data Figure 8: A)TRACERx and PEACE Cohort for ddPCR of 5 EGFRm. (i) Clinical information for each patient, (ii) Tumour EGFR mutation status, (iii) Normal EGFR mutation status. B) Representative H & E images from anthracotic pigment identification in TRACERx normal tissue. C) Comparing area of normal tissue harbouring anthracotic pigment in never smokers and smokers. Each dot represents the ratio of pigmented area respective to total tissue in each anthracosis positive normal lung tissue sample. D) Regression analysis of characteristics influences 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: Box plots summarise VAFs across samples. Mutations are grouped according to the gene harbouring the mutation and smoking status of the patient. BOTTOM: dot plots show VAFs of mutations in each sample. Where a sample has 2 mutations, they are both indicated. Dots are coloured by the gene harbouring the mutation (EGFR or KRAS) (Details of driver mutations can be found in Supplementary Table S5)

Methods

1. Normal Tissue Profiling

1.1) ddPCR of samples from TRACERx and PEACE studies

Tumour and normal lung tissue samples

This project leverages the infrastructure established by the national pan-cancer research autopsy programme (PEACE, NCT03004755) and the prospective, longitudinal cohort study (TRACERx) of non-small cell lung cancer (NCT01888601)[1].

To explore whether clinical disparities in never smoker lung cancer were reflected in normal lung tissue *EGFR* mutation status, we sought to assemble a cohort comprising TRACERx patients that were as best as possible balanced for sex (males vs females), smoking status (never smoker vs ever smoker) and *EGFR* mutation status in tumour samples (EGFRm vs EGFRwt). To uncover if *EGFR* mutations were also found in normal lung tissue from patients who never acquire a lung cancer diagnosis in their lifetimes, we also assembled a cohort of PEACE patients.

Based on tissue that was available for study, our dataset consisted of 195 tumour and 195 normal lung tissues from 195 TRACERx patients, and 59 normal lung tissues from 19 PEACE patients (median 3 samples per patient (range 1 to 10)).

In TRACERx, tumour and normal lung tissue were obtained at surgery. Normal lung tissue was collected distally from the primary tumour tissue (at least approximately 2cm apart). All tissue was initially snap-frozen and then a portion fixed and made into a FFPE block. A H&E section of each block was cut and stained and underwent pathology review. We use 'normal' to refer to non-malignant lung tissue. DNA was extracted from both the normal and tumor frozen tissue proximal to these sections. In PEACE, normal lung tissue was collected at post-mortem tissue harvest from patients who never acquire lung cancer in their lifetimes. Each piece of tissue collected was immediately bisected and one half snap frozen and the other fixed and then made into a FFPE block. H and E section of each block was cut and stained and underwent pathology review. DNA was then extracted from an adjacent normal frozen tissue sample.

All aforementioned H and E slides from tissues have undergone central pathology review. In particular, to exclude the possibility of contamination with tumour cells, thoracic pathologists have confirmed that all normal lung tissue samples do 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 is also noted.

EGFR mutation profiling in normal samples (with ddPCR)

DNA was extracted from normal lung tissue samples as previously described[1]. DNA concentration was measured using Qubit, and up to 3,000 ng of DNA was fragmented to approximately 1,500 bp using the Covaris E220 evolution Focused-ultrasonicator following the manufacturer's standard protocol. SAGAsafe assays[2] for 5 *EGFR* target variant alleles (EGFR L858R, *EGFR* Exon 19 del, *EGFR* S768I, *EGFR* L861Q and *EGFR* G719S) were employed (SAGA Diagnostics AB). SAGAsafe is a digital PCR-based ultra-sensitive mutation detection technology utilizing 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 rising from the 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 the Bio-Rad QX200 Droplet Digital PCR System. At least 3 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, 500ng of fragmented DNA (corresponding to ~150,000 copies of genome equivalents) was analyzed per assay across 4 reaction wells, with positive and negative control samples included in every run.

Calculation of copy number concentration of the variant and the wild-type alleles

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

Cvi 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 analyzed

Vd is the volume a droplet (0.85×10^{-3} μL)

Vr is the total volume of a ddPCR reaction (20 μL)

Vi is the input volume per ddPCR reaction of the input DNA sample

Calculation of the variant allele frequency (VAF)

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

Estimation of EGFRm rate

We considered all 5 oncogenic EGFR mutations detected via ddPCR in all TRACERx and PEACE (253 samples in total). Using the Approximate Bayesian computation model, we simulated ddPCR results of oncogenic EGFR mutations, and inferred a mutation rate of 4.07e-7 per mutation (confidence interval: 1.61e-7 to 6.08e-7). Considering this mutation rate, we estimated that the frequency of identifying 1 EGFRm (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). If we take the average of the 2 limits of the confidence interval, we obtain an estimate of an EGFRm being present in 1 in 554500 cells (or around 1:600,000 cells).

EGFR mutation profiling in corresponding tumour tissue (with MiSeq)

For each tumour region and matched germline, capture of a custom panel of genes (including the *EGFR* locus) was performed on 125ng DNA isolated from genomic libraries. The TruSeq Custom Amplicon Library Preparation method was used. Following cluster generation, samples were 100bp paired-end multiplex sequenced on the Illumina MiSeq at the GCLP lab at University College London, as described previously[1]. The generated data were aligned to the reference human genome (hg19) achieving a median sequencing depth of 3555X (Range: 1069-13084). Mutations were called as previously described[1].

1.2) DuplexSeq of samples from the PEACE and BDRE studies

Normal lung tissue samples

PEACE cohort samples are collected as described above. For DuplexSeq we obtained an additional normal lung tissue from 48 PEACE patients. Here, both lung cancer and other cancer type patients were profiled (lung cancer n=9; other cancer n=39)

All BDRE cohort patients were enrolled under Biomarker for Dysplastic Epithelium (BDRE) (NCT00900419). The cohort consisted of individuals recommended for 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 normal tissue and subsequently profiled for mutations using DuplexSeq. The absence of nodules or masses detected by chest CT scans was indicative of the non-tumor nature of 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 (with DuplexSeq)

Genomic DNA was extracted from brushings using Qiagen DNeasy Blood & Tissue kit according to manufacturer's instructions. Duplex libraries were prepared using a commercially available kit from TwinStrand Biosciences, Inc. (Seattle, WA, USA) (CKD-00042 panel 000323), starting with 250ng 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, DuplexSeq allows for the detection of rare mutations[3,4] with a sensitivity of less than 1 in 10^7. After shearing and capturing of gDNA spanning the panel, primers are ligated that allow the two strands of DNA for each segment to be uniquely labelled and matched with its opposing strand. These strands are then amplified and libraries were sequenced on the NovaSeq 6000 Sequencing System (Illumina Inc. San Diego, CA, USA) and sequencing data were analyzed on the DNAnexus platform. Samples had an average number of 150,000,000 raw reads, yielding a mean on-target duplex depth of 4500. DuplexSeq reads were processed using an in-house pipeline adapted from Valentine et al[5] and 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.

Data Availability

The MiSeq from the TRACERx and PEACE studies generated, used or analysed during this study are not publicly available and restrictions apply to the availability of these data. Such MiSeq data are available through the Cancer Research UK & University College London Cancer Trials Centre (ctc.tracerx@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 TRACERx data access committee, entering into an appropriate data access agreement and subject to any applicable ethical approvals.

The DuplexSeq data for the BDRE study were generated using a larger panel of probes that covered ~50 kb of the genome, spanning hotspots frequently mutated in cancers. All of the data for the EGFR and KRAS regions queried are included in this manuscript. Data for the other regions are not publicly available and restrictions apply to the availability of these data. Such DuplexSeq data are available through Professor James 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.

2. Epidemiological Studies

Study populations

2.1) UK Biobank dataset

Available Data

The UK Biobank (UKBB) study comprises over 500,000 participants, aged between 37-73 who were recruited between 2006-2010. Participants provide detailed information regarding a comprehensive set of lifestyle factors, in addition to physical measurements and biological samples. Particulate matter air pollution levels (in 2010) are estimated for addresses within 400km of the Greater London monitoring area using a land-use regression model developed as part of the ESCAPE study[6].

Associations between $PM_{2.5}$ and lung cancer incidence in the UKBB data have already been calculated and described in[7].

Imputing Missing Data

We first excluded all participants who had any cancer diagnosis pre-recruitment, alongside those with missing particulate matter or genetic principal components data. Multiple imputation with chained equations[8] was used to impute missing smoking status (categorised into “never”, “previous”, and “current”; <1% missing), passive smoking (weekly hours of home tobacco exposure; 10.0% missing), pack-years of smoking(15.4% missing), BMI (<1% missing), household income (dichotomised by >= £31,000 annually; 14.6% missing), educational attainment (split by degree/professional qualification status; 1.31% missing) values. In addition to these, imputation models also used the following variables to predict values for missing data: $PM_{2.5}$, age at baseline, sex, BMI, the first 15 genetic principal components (to account for ethnicity), 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 data set. This yielded 15 complete versions of the original dataset in which the missing values have been imputed. This data set comprised 407,509 individuals and represented 27 cancer types. Each imputed dataset was independently used in the same analysis protocol.

Cox Regression To Identify Associations Between $PM_{2.5}$ & Cancer Incidence

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 one sex), and pooled results across these models, which were consistent for each cancer type, into a single set using Rubin’s rules[8]. Confidence intervals were calculated using: $e^{\,estimate_{pooled} \pm (1.96*standard\ error_{pooled})}$. 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 due 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 principal component 12), kidney (age and smoking status), and melanoma (genetic principal component 9 and sex). We note high median (across all 15 imputations) variance inflation factor values (VIF >= 5) for the following covariates: genetic principal component 1 (other and unspecified biliary tract parts), 2 (AML, follicular nodular NHL, larynx, mesothelioma, other and unspecified biliary tract parts, peripheral and cutaneous T lymphomas, retroperitoneum and peritoneum), and 3 (AML, follicular nodular NHL, larynx, mesothelioma, other and unspecified biliary tract parts, peripheral and cutaneous T lymphomas). Finally, we report FDR-corrected p-values for the $PM_{2.5}$-cancer incidence association, to account for multiple testing.

Our methods differed from those of Huang et al., in the following ways:

- We made use of the more refined cancer registry data rather than hospital diagnosis data. In particular, some of the cancer type definitions for lung and renal cancers are refined and updated in Supplementary Table S1. We also changed the censoring date to the last day of 2018 instead of it being the date of last diagnosis of each cancer type.
- To enhance the robustness of our work:
    - We excluded $PM_{2.5-10}$ and PM_{10} (due to collinearity) and participants with any cancer diagnosed pre-baseline
    - Used age at baseline instead of age at diagnosis
    - Included additional dependent variables (cancer diagnosis and time till end of follow-up) in imputation models
    - Increased the number of imputations from 5 to 15 and iterations from 90 to 180
    - Augmented our multivariate analysis to better account for the effect of smoking by categorising participants into "never", "previous", and "current" smokers. We also controlled for the smoking intensity by including pack-years of smoking as a continuous variable in our regression models.

Interaction test between $PM_{2.5}$ and smoking

An interaction test between $PM_{2.5}$ and smoking was performed for lung cancer, considering only participants with complete covariate data in the multivariable Cox regression.

LUAD-specific analysis

We considered only participants with cancer registry histology entries that map to LUAD (Supplementary Table S1). Imputations and all downstream modelling was performed independently for this analysis.

Analysis taking into account migration

Since the $PM_{2.5}$ data is available for each participant's address, we assume that participant $PM_{2.5}$ exposures remain constant throughout the study period. To account for exposure miss-classification, we additionally performed a separate analysis including only participants who had lived at their current address for at least three years prior to baseline. All imputations and downstream analysis was performed independently for this subgroup.

Association between radon exposure and lung cancer incidence

Radon exposure data from PHE was merged with the UKBB dataset based on home location coordinates. Since the data from PHE had greater spatial resolution, values were aggregated by the mode radon potential class (breaking ties through taking the higher class value) across all PHE 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 $PM_{2.5}$.

2.1.1 Comparison of UKBB Population with General UK Population

We have provided a table (Supplementary Table S4) 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, more likely to be female, older, White, and live in areas with lower $PM_{2.5}$ levels.

2.2) Within-country datasets

2.2.1) England dataset (Public Health England)

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.

Air pollution: Annual $PM_{2.5}$ air pollution data (μg/m^3) from 2006 to 2017 was obtained at the grid code level (1km x 1km) from DEFRA[9]. 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 code level (UK Health Security Agency (UKHSA)-British Geological Survey (BGS). Radon data: indicative atlas of radon in Great Britain. https://www.bgs.ac.uk/datasets/radon-data-indicative-atlas-of-radon/). Postal code coordinates were sourced from the ONS 2018 Postal Code Directory[10]. 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 then aggregated to the Cancer Alliance region level and then averaged over the period 2008 to 2017 for 2018 diagnoses, 2007 to 2016 for 2017 diagnoses and 2006 to 2018 for 2016 diagnoses - these were selected because they represented the 10 years prior to 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.

Lung cancer incidence: Data on 118,019 (2016: 39,229, 2017: 39,500, 2018: 39,290) lung cancers (International Classification of Diseases codes C33 to C34) diagnosed in England between 1 January 2016 and 31 December 2018 were extracted from the National Cancer Registration Dataset (NCRD) [AV2018 in CASREF01 (end of year snapshot)], held by the National Disease Registration and Analysis Service at Public Health England. Lung cancer incidence for each Cancer Alliance region was calculated based on these cases. This represented a predominantly Caucasian cohort - White: 92.06%, Asian: 1.48%, Chinese: 0.23%, Black: 1.05%, Mixed: 0.28%, Other: 0.94%, Unknown: 3.96%.

The age-standardised lung cancer incidence (using population counts obtained from the Office of National Statistics 2019 (2018 mid-year estimates)) was obtained according to each five-year age group and sex. Incidences were then combined across age and sex to yield a single value for each alliance region.

Lung cancer incidence = (sum(wi*xi/di)/sum(wi)) * 100000
  wi = European population standard
  di = Population Count
  xi = Case Count

Standardised rates are standardised according to the 2013 European Standard Population. Confidence intervals for ASR point estimates were calculated using the Dobson method.

EGFR mutation proportion: For lung cancer diagnoses listed above, *EGFR* mutation statuses were extracted from the NCRD [AT_GENE_ENGLAND table in the CAS2210 monthly snapshot], which includes data on somatic tests undertaken from 1st January 2016 to 31st December 2019. Only cases with “Overall: TS” as “a:abnormal” and “b:normal” for EGFR were used in the calculation for EGFR mutation rate (n=25,567). The EGFR mutation rate was calculated for each Cancer Alliance region.

EGFR mutation rate =<# a:abnormal> / (<# a:abnormal> + <# b:normal>)

2.2.2) 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.

Air pollution: $PM_{2.5}$ air pollution data were obtained from Air Korea[11] for the years 2015 to 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 prior to the year of lung cancer diagnosis. $PM_{2.5}$ levels between 2015 to 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 Korean region was only available starting from 2015.

Lung cancer incidence: Lung cancer incidence data were obtained from the Korean National Cancer Center[12] 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.

EGFR mutation proportion: 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=2563). EGFR mutation rate was calculated as above.

2.2.3) 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.

Air pollution: Annual $PM_{2.5}$ air pollution data was obtained for 12 standard geographical regions in Taiwan from the Environmental Protection Administration Executive Yuan R.O.C. (Taiwan)[13]. $PM_{2.5}$ (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) prior to the year of lung cancer diagnosis. Eg. For a diagnosis in 2017, 2006-2015 aggregated air pollution levels were used for analysis; while 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 dramatic decreases in air pollution levels after 2013.

Lung cancer incidence: 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=4599). 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 (CGMH), and the age-standardized rates (ASR) per 100,000 were calculated using the world (WHO 2000) standard population of lung cancer incidence.

EGFR mutation proportion: *EGFR* mutation testing data were available for all of these cases. However, only 9 counties had at least 10 cases with EGFR mutation tested per year and comprised >5% of the total population, these were the counties that were retained for analysis. EGFR mutation rate was calculated as above.

Relationship between EGFRm lung cancer incidence and $PM_{2.5}$

Analyses were performed separately for each of the 3 cohorts: England, South Korea, and Taiwan.

For each geographical region (eg. each country; the 20 cancer alliances in England), *EGFR* mutant 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).

EGFRm lung cancer incidence = <lung cancer incidence>*<EGFR mutation rate>

EGFR mutant lung cancer incidence values were compared with mean $PM_{2.5}$ values across geographical regions using:

1. Pearson correlation tests
2. Weighted Pearson correlation tests (to account for number of tested cases in each geographical region)

3. Robust linear regression (to account for outliers)

Sensitivity analysis for England and Korea data sets

In the England data set, there were 2 Cancer Alliance regions (South East London and Thames Valley) with sparse data due to data unavailability (<5% of lung tumours have any molecular testing data recorded (2016-2018)). To exclude the possibility of this confounding our analysis, we performed a sensitivity analysis, where we excluded data from these 2 regions. Of note, the correlation between $PM_{2.5}$ and EGFRm lung cancer incidence was still significant (R=0.55; p=0.019) after these exclusions.

Similarly, in the South Korea data set Jeju-do (2017) was excluded due to poor data availability. The correlation between $PM_{2.5}$ and EGFRm lung cancer incidence was still significant (R=0.38; p=0.033) after this exclusion.

However, for the sake of completion, we have reported the full data sets (including these 2 England regions and 1 South Korea region) in the main text.

2.3) Canada Data Set (BC Cancer Research Centre, Vancouver BC, Canada)

This data set comprises 228 female lung cancer cases that have been reported in Myers et al 2021[14]. These cases were seen at the Thoracic Surgery Department of the Vancouver General Hospital or the BC Cancer Vancouver Cancer Center between November 15, 2017, and May 31, 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 allow accurate linking of residential locations to satellite-derived $PM_{2.5}$ exposure data that were available from 1996 onward. A personal PM 2.5 cumulative exposure was individually calculated via a detailed residential history from birth to current address, and input into geographical information System mapping (GIS). By applying high resolution (10X10 km) concentration estimates of particulate matter <2.5um from satellite observations, chemical transport models and ground measurements to each individual’s residential history, a cumulative exposure 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.

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 whole data set (3 year cumulative pollution data and 20 year cumulative pollution data).

Thresholds
Bottom quintile: 6.77ug/m3
Top quintile: 7.27ug/m3

$PM_{2.5}$ Low: $PM_{2.5}$<bottom quintile
$PM_{2.5}$ Intermediate: $PM_{2.5}$>bottom quintile & $PM_{2.5}$<top quintile
$PM_{2.5}$ High: $PM_{2.5}$>top quintile

Comparing EGFRm frequencies

EGFRm frequencies were compared between high and low pollution exposure groups using Chi-squared tests. 2 comparisons were performed:

- High vs Low Pollution (based on 3yr data)
- High vs Low Pollution (based on 20 yr data)

3. 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.

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 Jackson laboratory. Mice were backcrossed onto a C57Bl6/J background and further crossed to generate Rosa26$^{\text{LSL-tTa/LSL-tdTomato}}$/Tet(O)*EGFR*$^{\text{L858R}}$ mice. Rosa26rtTa/TetO-*EGFR*$^{\text{L858R}}$ and Rosa26$^{\text{LSL-tdTomato}}$;LSL-*Kras*$^{\text{G12D}}$ mice have been described previously[15,16] . After weaning, the mice were genotyped (Transnetyx, Memphis, USA), and placed in groups of one to five animals in individually ventilated cages, with a 12-hour daylight cycle. Cre-mediated recombination was initiated by adenoviral CMV-Cre (Viral Vector Core, University of Iowa, USA) delivered via intratracheal intubation (2.5×10^{7} virus particles/50 µl) or by Ad5-SPC-Cre delivery (Viral Vector Core, University of Iowa, USA**,** donated by Dr. Anton Berns from the Netherlands Cancer Institute) delivered via intratracheal instillation (2.5×10^{8} virus particles/50 µl**).**

For exposure to fine particulate matter or control, SRM2786 from the National Institute of Standards and Technologies (NIST, obtained from Sigma Aldrich) was resuspended in sterile PBS using sonication and 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 (Schantz et al., 2016). Mice were briefly anesthetized using 5% isoflurane and intratracheal administration of 5 µg, 50 µg or control PBS was performed.

Fluorescence-activated cell sorting analysis and cell sorting

For flow cytometry analysis of immune cells, mouse lungs were minced into small pieces, incubated with collagenase (1 mg/ml; ThermoFisher) and DNase I (50 U/ml; Life Technologies) for 45 min at 37°C and filtered through 100 μm strainers (Falcon). Red blood cells were lysed for 5 min using ACK buffer (Life Technologies). Cells were stained with fixable viability dye eFluor870 (BD Horizon) for 30 min and blocked with CD16/32 antibody (Biolegend) for 10 min. Cells were then stained with antibodies for 30 min (see Supplementary Table S6). Intracellular staining was performed using the Fixation/Permeabilization kit (eBioscience) according to the manufacturer's instructions. Samples were resuspended in FACS buffer (2% fetal calf serum in PBS) and analysed using a BD Symphony flow cytometer. Data was analysed using FlowJo (Tree Star).

For flow cytometry sorting of epithelial 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 (300 x g, 5 min, 4 degrees and red blood cells lysed as above. Extracellular antibody staining was then performed followed by incubation in DAPI (Sigma Aldrich) to label dead cells. Sorting strategies are outlined in Extended Data 6B,C. Cell sorting was performed on Influx, Aria Fusion or Aria III machines (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 the following antibodies human EGFR L858R mutant specific (Cell Signaling: 3197, 43B2), anti-RFP (Rockland: 600-401-379), CD11b (ab133357) and CD68 (ab283654). Primary antibodies were detected either using biotinylated secondary antibodies, followed by HRP/DAB or with subsequent OPAL fluorescence secondary antibodies (Akoya) . A commercial kit was used to detect IL1b RNA transcripts by RNAscope (ACD Biotechne) following manufacturers instructions, and staining for CD68 protein was performed afterwards and detected using OPAL fluorescence following manufacturers protocols (Akoya). And probes visualised using fluorescence to detect IL1b RNA and CD68 protein simultaneously. Slides were imaged using a Leica Zeiss AxioScan.Z1 slide scanner.

Tumour grading was carried out by two board-certified veterinary pathologists. Tumour 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[17]. 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, based on the low proliferation rate of alveolar epithelial cells (Desai et al.

2014). Since there was an average cluster size of 2 EGFR mutant cells at 3 weeks, we defined clusters >5 cells at 10 weeks as 'expanded clusters' that grew above expected

Whole Genome Sequencing (WGS)

Lung tumours from control-treated mice (PBS) (n=5) and PM exposed mice(n=5) were collected at ethical endpoint. Individual lung tumours were dissected from lung lobes and snap frozen. Germline DNA was extracted from tail tissue. DNA was isolated and prepared for WGS, followed by sequencing on a NovaSeq (Ilumina), to achieve target coverage of 100X for PBS and PM exposed samples, and 30X for germline samples. Sequences from all 20 samples were processed using the Nextflow (version 21.10.3) Sarek pipeline (nf-core/sarek v3.0). Briefly, sequences were aligned with BWA (0.7.17) to mm10, and mutations were called with Mutect2 (gatk4: 4.1.8.1). Only passed mutations that were uniquely present in each tumour were considered for analysis. Mutational signatures were called using the DeconstructSigs R package, restricting our analysis to the common single base substitution signatures: SBS1, SBS4, SBS5, SBS2, SBS13, SBS40, SBS92, SBS17a, SBS17b, SBS18.

Driver mutation probability

The list of driver mutations and the mutational signature exposures are obtained from the TRACERx 421 publication[18]. Only patients with detected smoking-related signatures are considered in the analysis (TRACERx 421). Each observed clonal driver mutation is given a probability to be caused by all active mutational signatures in the patient. This number is derived from multiplying the mutational signatures exposures to the 96-channel profile of each signature[19]. Then the value is normalised to 1, so that each driver mutation can be explained by a fraction of active mutational signatures. The probabilities are then aggregated, giving the overall contribution to driver mutations from each of the active mutational signatures. A patient is defined as non-carrier of a tobacco-related driver mutation if the probability of SBS4 and SBS92 (smoking-related signatures) is smaller than 0.5.

RNA-Sequencing (RNA-seq)

Lung CD45− CD31− Ter119− EpCAM+ were sorted from control and PM exposed mice by flow cytometry. Total RNA was isolated using the miRNeasy Micro Kit (Qiagen), according to the manufacturer's instructions. Library generation was performed using the KAPA RNA HyperPrep with RiboErase (Roche), followed by sequencing on a HiSeq (Ilumina), to achieve an average of 25 million reads per sample.

The RNA-seq pipeline of nf-core framework version 3.3 was launched with Nextflow version 21.04.0 to analyse RNA sequencing data[20]. Raw reads in fastq files were mapped to GRCm38 with associated ensemble transcript definitions using STAR version 2.7.6a[21]. BAM files were sorted with a chromosome coordinate using samtools version 1.12 . RSEM version 1.3.1 was used to calculate estimated read counts per gene and to quantify in a measure of transcripts per million (TPM)[22].

Differential expression analysis was performed using the R platform version 4.0.3 package DESeq. filtering with the absolute value of log fold change more 1 and p-value less than 0.05[23]. The gene expression between treatment groups was further analysed for their pathway enrichments using Gene Set Enrichment Analysis (GSEA). Normalisation (using z-scores) of TPM scores across the dataset was performed prior to plotting heatmaps of gene expression.

The AT2-like score was derived using the method described by Young et al[24]. Briefly, bulk RNA-seq data from mouse models, with or without an EGFR mutation and in the presence or absence of PM, were compared according to the degree to which they were similar to a signature of keratin8+ AT2 transitional stem cells derived from single cell RNAseq data from Strunz et al.[25]. Gene expression within genes overlapping the human and mouse genomes was used as input, and the pseudoR2 value from the Young et al approach used as a continuous variable in a test between the different conditions.

Comparison of RNA-seq data from mice to never-smokers in COPA study

RNA sequencing was applied to 18 samples of bronchial brushings from 9 never-smokers from the COPA study after exposure to filtered air and diesel exhaust. Salmon[26] was used to estimate transcript-level abundance from RNA-seq read data. Differential expression analysis was performed using DESeq2[27]. The log two fold difference in gene expression was calculated between samples collected 24 hours 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 two fold change of significantly differentially expressed genes between the tdTomato control and tdTomato PM-treated mice were compared to the log two fold change expression of the genes from COPA participants.

The limitation of this analysis is that the mouse and human RNA-seq datasets fundamentally differ:

- Mouse data is acquired from total lung EpCAM+ cells, containing both airway and alveolar tissue, whilst the human data were obtained from bronchial brushings only, hence different cell types are represented in the data.
- The pollution exposure between species differed; human participants were exposed to diesel exhaust for 2 hours, compared to 3 weeks of PM exposure for mice. Furthermore, the mice were kept in controlled environments, whereas a 4 week wash-out period between exposure to filtered air and diesel exhaust in human participants was required, where day-to-day particulate matter exposures and lifestyle differences could not be controlled.
- Fold changes from the human data were obtained by pairwise comparisons from each individual, while since 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 (ie. air pollution vs control).

- As well, the RNA-seq was performed at 2 different sequencing centres, and target depths were different. The human data was sequenced with a target depth of 30M reads/sample, while the mouse data was sequenced with a target depth of 25M reads/sample.

Organoid forming assays

Lung organoid co-culture assays have been previously described[28]. Briefly, tdTomato+ lung epithelial cells (tdTomato+EpCAM+CD45− CD31− Ter119−) were isolated by fluorescence-activated cell sorting (FACS) from control or PM exposed ET mice acutely after 3 weeks of treatment and were resuspended in 3D organoid media consisting of DMEM/F12 with 10% FBS, 100 U ml− 1 penicillin-streptomycin, insulin/transferrin/selenium, L-glutamine (all GIBCO) and 1mM HEPES (in-house). 5,000-10,000 cells were mixed with a murine lung fibroblast cell line (MLg2908, ATCC, 1:5 ratio) and resuspended in growth factor reduced Matrigel (Corning) at a ratio of 1:1. 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 °C, 500 µl of organoid media was added to the lower chamber and media changed every other day. Bright-field and fluorescent images were acquired after 14 days using an EVOS microscope (Thermo Fisher Scientific) and quantified using FiJi (.2.0.0-rc-69/1.52r, ImageJ).

For *ex vivo* interleukin-1-beta treatment of lung alveolar type II (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−)[24]. Purified AT2 cells were incubated in vitro with 6 x 10^7 PFU/ml of Ad5-CMV-Cre in 100uL per 100,000 cells 3D organoid media for 1hr at 37 C as detailed in[30]. Cells were washed three times in PBS before plating as above, with 20ng/mL IL-1b added to the organoid media in the lower chamber and changed every other day. TdTomato+ organoids were counted as above and the size analysed in FiJi. For wholemount staining of organoids, organoids were prepared according to previous methods[31] and stained with anti-proSPC (Abcam, clone EPR19839) and anti-keratin 8 (DSHB Iowa, clone TROMA-1). 3D confocal images were acquired upon an Olympus FV3000 and analysed in FiJI.

For assessment of AT2 organoid formation after PM exposure, AT2 cells were isolated from control or PM treated T and ET mice after 3 weeks, without *in vivo* Cre induction, followed by Cre infection as above and 10,000 cells plated in the organoid assay as described. 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 isolation of both AT2 cells, interstitial and alveolar macrophages as detailed in Choi et al[32], (sorting strategies defined in Extended Data Figure 6C). AT2 cells from PBS-treated ET mice only were infected with Cre ex vivo as described as above, before 10,000 AT2 cells were either plated with fibroblasts only, or with a 1:6 ratio of PBS- or PM- treated macrophages as above, modified from Choi et al. tdTomato+ organoids were quantified in all conditions.

Statistics and Reproducibility

Preclinical statistical analyses were performed using Prism (v.9.1.1, GraphPad Software). Epidemiological and mutation/sequence data analysis was performed in R version 3.6.2. Graphic display was performed in Prism and illustrative figures created with Biorender.com. A Kolmogorov–Smirnov normality test was performed before any other statistical test. After, if any of the comparative groups failed normality (or the number 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 one-way analysis of variance (ANOVA) or nonparametric Kruskal–Wallis test.

No data were excluded. No statistical methods were used to predetermine sample size in the mouse studies, and 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.

Code Availability

Normal lung tissue processing, RNA-seq analysis and WGS analysis code available at: https://github.com/emilialim/airpoll_cancer

Online Methods References

1. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
2. Dahlgren, M. *et al.* Preexisting Somatic Mutations of Estrogen Receptor Alpha (ESR1) in Early-Stage Primary Breast Cancer. *JNCI Cancer Spectr.* **5**, pkab028 (2021).
3. Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
4. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14508–14513 (2012).
5. Valentine, C. C. *et al.* Direct quantification of in vivo mutagenesis and carcinogenesis using duplex sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33414–33425 (2020).
6. Eeftens, M. *et al.* Development of Land Use Regression models for PM(2.5), PM(2.5)

absorbance, PM(10) and PM(coarse) in 20 European study areas; results of the ESCAPE project. *Environ. Sci. Technol.* **46**, 11195–11205 (2012).

7. 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.* **204**, 817–825 (2021).

8. Buuren, S. van & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).

9. Department for Environment, F. and R. A. (Defra) webmaster@defra gsi gov uk. Modelled background pollution data- Defra, UK. https://uk-air.defra.gov.uk/data/pcm-data#population_weighted_annual_mean_pm25_data.

10. ONS Postcode Directory (Latest) Centroids. https://geoportal.statistics.gov.uk/datasets/ons-postcode-directory-latest-centroids/explore?showTable=true.

11. 에어코리아. https://www.airkorea.or.kr/web.

12. 암등록통계자료 > 중앙암등록본부 > 국가암관리사업 | 국립암센터. https://ncc.re.kr/cancerStatsList.ncc?sea.

13. Administration, E. P. & 行政院環境保護署. Environmental Protection Administration – Taiwan Air Quality Monitoring Network. https://airtw.epa.gov.tw/ENG/Default.aspx.

14. Myers, R. *et al.* High Ambient Air Pollution Exposure Among Never Smokers Versus Ever Smokers with Lung Cancer. *J. Thorac. Oncol.* (2021) doi:10.1016/j.jtho.2021.06.015.

15. Politi, K. *et al.* Lung adenocarcinomas induced in mice by mutant EGF receptors found in human lung cancers respond to a tyrosine kinase inhibitor or to down-regulation of the receptors. *Genes Dev.* **20**, 1496–1510 (2006).

16. Jackson, E. L. *et al.* Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. *Genes Dev.* **15**, 3243–3248 (2001).

17. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

18. Frankell, A., Dietzen, M., Al Bakir, M. & Lim, E. The evolution of lung cancer and impact of subclonal selection in TRACERx. *Nature* **In Press**,.

19. Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In silico saturation mutagenesis of cancer genes. *Nature* **596**, 428–432 (2021).

20. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).

21. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

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

23. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

24. Young, M. D. *et al.* Single cell derived mRNA signals across human kidney tumors. *Nat. Commun.* **12**, 3896 (2021).

25. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nat. Commun.* **11**, 3559 (2020).

26. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

27. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

28. Nolan, E. *et al.* Radiation exposure elicits a neutrophil-driven response in healthy lung tissue that enhances metastatic colonization. *Nat. Cancer* **3**, 173–187 (2022).

29. Major, J. *et al.* Type I and III interferons disrupt lung epithelial repair during recovery from viral infection. *Science* **369**, 712–717 (2020).

30. Dost, A. F. M. *et al.* Organoids Model Transcriptional Hallmarks of Oncogenic KRAS

Activation in Lung Epithelial Progenitor Cells. *Cell Stem Cell* **27**, 663-678.e8 (2020).

31. Dekkers, J. F. *et al.* Long-term culture, genetic manipulation and xenotransplantation of human normal and breast cancer organoids. *Nat. Protoc.* **16**, 1936–1965 (2021).

32. Choi, J. *et al.* Inflammatory Signals Induce AT2 Cell-Derived Damage-Associated Transient Progenitors that Mediate Alveolar Regeneration. *Cell Stem Cell* **27**, 366-382.e7 (2020).

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number [82693]. This work has been supported by the Mark Foundation ASPIRE I Award (Grant 21-029-ASP), Lung Cancer Research Foundation Grant on Disparities in Lung Cancer, Advanced Grant (PROTEUS, Grant Agreement no. 835297), CRUK EDD (EDDPMA-Nov21\100034), and Rosetrees Out-of-round Award (OoR2020\100009). E.L.L. receives funding from NovoNordisk Foundation (ID 16584), The Mark Foundation (Grant 21-029-ASP) and has been supported by Rosetrees. 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. C.E.W. is supported by a RESPIRE4 fellowship from the European Respiratory Society and Marie-Sklodowska-Curie Actions. K.C. is supported by Research Unit of Intelligence Diagnosis and Treatment in Early Non-small Cell Lung Cancer, Chinese Academy of Medical Sciences (2021RU002), 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 of 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). 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 also receives funding from Cancer Research UK, Rosetrees and the NIHR BRC at University College London Hospitals and the CRUK University College London Experimental Cancer Medicine Centre. EJE is supported by the Blumenthal Fellowship from the Linda Crnic Institute for Down Syndrome. J.D., M.G., Y.E.M. D.T.M. and R.L.K receive funding from American Association for Cancer Research/Johnson&Johnson (18-90-52-DEGR), and J.D. is supported by the Courtenay C. and Lucy Patten Davis Endowed Chair in Lung Cancer Research. M.G., Y.E.M. D.T.M. and R.L.K. were supported by National Cancer Institute (NCI) RO1 CA219893. E.J.E. was supported by NCI Ruth L. Kirschstein National Research Service Award T32-CA190216. The work at the University of Colorado was also supported by NCI Cancer Center Support Grant P30CA046934. M.J.-H. has received funding from Cancer Research UK, National Institute for Health Research, Rosetrees Trust, UKI NETs and NIHR University College London Hospitals Biomedical Research Centre. C.S. is Royal Society Napier Research Professor. He is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001169), the UK Medical Research Council (FC001169), and the Wellcome Trust (FC001169). C.S. is funded by Cancer Research UK (TRACERx, PEACE and CRUK Cancer Immunotherapy Catalyst Network), Cancer Research UK Lung Cancer Centre of Excellence, the Rosetrees Trust, Butterfield and Stoneygate Trusts, NovoNordisk Foundation (ID16584), Royal Society Research Professorships Enhancement Award (RP/EA/180007), the NIHR BRC at University College London Hospitals, the CRUK-UCL Centre, Experimental Cancer Medicine Centre and the Breast Cancer Research Foundation (BCRF). This research is supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (SU2C-AACR-DT23-17). Stand Up To Cancer is a program of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. C.S. also receives funding from the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013) Consolidator Grant (FP7-THESEUS-617844), European Commission ITN (FP7-PloidyNet 607722), an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union’s Horizon 2020 research and innovation programme (835297) and Chromavision from the European

Union’s Horizon 2020 research and innovation programme (665233).This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (grant no. FC001112), the UK Medical Research Council (grant no. FC001112), and the Wellcome Trust (grant no. FC001112) and the European Research Council (grant no. ERC CoG-H2020-725492).

We thank the PEACE clinical and admin team for their assistance in data curation: Scott Shepherd, Zayd Tippu, Ben Shum, Charlotte Lewis, Molly O’Flaherty, Analyn Lucanas, Eleanor Carlyle, Lucy Holt, Fiona Williams. We also thank nursing and biospecimen coordinators for their assistance in sample curation: Kim Edmonds, Lauren Grostate, Karla Lingard, Denise Kelly, Justine Korteweg, Lauren Terry, Jennifer Biano, Aida Murra, Kayleigh Kelly, Kema Peat, Nikki Hunter. We thank Alvin Ho-Kwan Cheung for assistance in pathology review. We thank Johanna Asklin and Cecilia Forsberg for logistical and technical assistance, and the Chang Gung Memorial Hospital for providing Chang Gung Research Database (CGRD) data. We are also grateful for support from the Flow Cytometry Unit, the Experimental Histopathology Unit, the Advanced Light Microscopy Facility, the Advanced Sequencing Facility and the Biological Resources Unit, especially Nicholas Chisholm and Jay O’Brien. We thank Agnes Yuen, Ayman Azhar, Kevin Lau, Carley Schwartz, Andrew Lee, and Chris Rider for their logistical support for the human exposure study. We also thank Centre d’expertise et de services Génome Québec for their sequencing services and support

Author Contributions

W.H. and E.L.L jointly designed the project, performed the experiments analyses and wrote the manuscript. W.H. performed the mouse experiments, E.L.L. performed the bioinformatics and epidemiology analyses. C.E.W. performed the mouse experiments, helped to 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 UK Biobank analyses. K.C. assembled and analyzed the TRACERx cohort. F.-C.K. and M.-H.L. performed the Taiwan epidemiological analyses. F.M., E.J.E.J., C.T., M.G., Y.E.M., D.T.M., and R.L.K. generated and analyzed 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.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 to analyse patient clinical characteristics. D.M. 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.L., J.P., S.H., F.R. curated the PHE data set, 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.D., T.J., N.K. and E.G. provided supervision over mouse experiments. M-J. F., M.H., P.A. and N.M. guidance supervision over bioinformatics analyses. S.L., P.S., R.H., C.T., C.D.B., A.H. and K.L. provided supervision over epidemiological analyses. C.C. provided supervision over human cross over study. J.D.G. designed the BDRE study and supervised the normal tissue profiling work. M.J.-H. designed PEACE and TRACERx study protocols and E.L.L. and M.J.-H. jointly supervised the study and collaborations. C.S. designed and supervised the study and helped to write the manuscript.

Competing Interests

M.A.B. has consulted for Achilles Therapeutics. 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 of Dragonfly Therapeutics, SQZ Biotech and Skyhawk Therapeutics. T.J. is also President of Break Through Cancer. M.J.-H. is a CRUK Career Establishment Awardee and has received funding from CRUK, IASLC International Lung Cancer Foundation, Lung Cancer Research Foundation, Rosetrees Trust, UKI NETs, NIHR, NIHR UCLH Biomedical Research Centre; has consulted and is a member of the Scientific Advisory Board and Steering Committee for Achilles Therapeutics; has received speaker honoraria from Astex Pharmaceuticals, and Oslo Cancer Clusters; and holds a patent to methods for lung cancer detection (PCT/US2017/028013). N.M. has stock options in and has consulted for Achilles Therapeutics and holds a European patent in determining HLA LOH (PCT/GB2018/052004), and is a co-inventor to a patent to identifying responders to cancer treatment (PCT/GB2018/051912). C.D.B. has partnerships with GRAIL, LLC, NHS Galleri Trial, IDMC, Mercy BioAnalytics, Lucid DX,Medial EarlySign. C.S. acknowledges grant support from Pfizer, AstraZeneca, Bristol Myers Squibb, Roche-Ventana, Boehringer-Ingelheim, Archer Dx Inc. (collaboration in minimal residual disease sequencing technologies) and Ono Pharmaceutical; is an AstraZeneca Advisory Board Member and Chief Investigator for the MeRmaiD1 clinical trial; has consulted for Amgen, Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol Myers Squibb, AstraZeneca, Illumina, Genentech, Roche-Ventana, GRAIL, Medicxi, Bicycle Therapeutics, Metabomed and the Sarah Cannon Research Institute; has stock options in Apogen Biotechnologies, Epic Bioscience and GRAIL; and has stock options and is co-founder of Achilles Therapeutics. C.S. holds patents relating to assay technology to detect tumour recurrence (PCT/GB2017/053289); to targeting neoantigens (PCT/EP2016/059401), identifying patent response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221), to treating cancer by targeting Insertion/deletion mutations (PCT/GB2018/051893), identifying insertion/deletion mutation targets (PCT/GB2018/051892); methods for lung cancer detection (PCT/US2017/028013), identifying responders to cancer treatment (PCT/GB2018/051912); and a patent application to determine methods and systems for tumour monitoring (GB2114434.0).