Non-Small-Cell Lung Cancer Promotion by Air Pollutants

Charles Swanton (charles.swanton@crick.ac.uk)
The Francis Crick Institute https://orcid.org/0000-0002-4299-3018

William Hill
The Francis Crick Institute

Emilia Lim
The Francis Crick Institute

Claudia Lee
The Francis Crick Institute

Clare Weeden
The Francis Crick Institute

Marcellus Augustine
Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK.

Kezhong Chen
Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK.

Feng-Che Kuan
Department of Hematology and Oncology, Chang Gung Memorial Hospital, Chiayi Branch, Chiayi, Taiwan.

Fabio Marongiu
Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, Colorado.

Edward Evans
Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, Colorado. https://orcid.org/0000-0001-8120-3755

David Moore
University College London

Felipe Rodrigues
The Francis Crick Institute

Febe van Maldegem
The Francis Crick Institute; Department of Molecular Cell Biology and Immunology, Amsterdam UMC, Location VUMC, Amsterdam, The Netherlands

Jesse Boumelha

The Francis Crick Institute

Selvaraju Veeriah

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

Andrew Rowan

Translational Cancer Therapeutics Laboratory, Francis Crick Institute

Cristina Naceur-Lombardelli

UCL Cancer Institute

Takahiro Karasaki

Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute https://orcid.org/0000-0001-6863-2360

Monica Sivakumar

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

Deborah Caswell

The Francis Crick Institute

Ai Nagano

The Francis Crick Institute

Min Hyung Ryu

Department of Medicine, Division of Respiratory Medicine, Chan-Yeung Centre for Occupational and Environmental Respiratory Disease, Vancouver Coastal Health Research Institute, UBC, Vancouver, BC, C

Ryan Huff

Department of Medicine, Division of Respiratory Medicine, Chan-Yeung Centre for Occupational and Environmental Respiratory Disease, Vancouver Coastal Health Research Institute, UBC, Vancouver, BC, C

Shijia Li

Department of Medicine, Division of Respiratory Medicine, Chan-Yeung Centre for Occupational and Environmental Respiratory Disease, Vancouver Coastal Health Research Institute, UBC, Vancouver, BC, C

Alastair Magness

The Francis Crick Institute

Alejandro Suarez-Bonnet

Royal Veterinary College https://orcid.org/0000-0003-0296-5896

Simon Priestnall

Royal Veterinary College

Margreet Lüchtenborg

National Disease Registration Service, Public Health England, Wellington House, London, UK

Katrina Lavelle

National Disease Registration Service, Public Health England, Wellington House, London, UK

Joanna Pethick

National Disease Registration Service, Public Health England, Wellington House, London, UK

Steven Hardy

National Disease Registration Service, Public Health England, Wellington House, London, UK

Fiona McRonald
National Disease Registration Service, Public Health England, Wellington House, London, UK
Meng-Hung Lin
Chang-Gung Memorial Hospital Chiayi Branch https://orcid.org/0000-0002-8594-3360
Clara Troccoli
Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, Colorado.
Moumita Ghosh
Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine.
York Miller
Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine;Veterans Affairs Eastern Colorado Healthcare System, Aurora, Colorado.
Daniel Merrick
Department of Pathology, University of Colorado Anschutz Medical Campus, Aurora, Colorado.
Robert Keith
Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine.
Maise Al Bakir
CRUK Lung Cancer Centre of Excellence; The Francis Crick Institute
Chris Bailey
The Francis Crick Institute
Lao Saal
SAGA Diagnostics AB, Lund, Sweden
Yilun Chen
SAGA Diagnostics AB, Lund, Sweden
Anthony George
SAGA Diagnostics AB, Lund, Sweden https://orcid.org/0000-0001-5680-8664
Chris Abbosh
UCL https://orcid.org/0000-0002-8983-1382
Nnennaya Kanu
The Francis Crick Institute
Se-Hoon Lee
Sungkyunkwan University School of Medicine
Nicholas McGranahan
University College London
Chistine Berg
National Cancer Institute, USA (retired)
Eva Grönroos
The Francis Crick Institute https://orcid.org/0000-0001-8303-5409
Julian Downward

The Francis Crick Institute https://orcid.org/0000-0002-2331-4729

Tyler Jacks

David H. Koch Institute for Integrative Cancer Research https://orcid.org/0000-0001-5785-8911

Christopher Carlsten

Department of Medicine, Division of Respiratory Medicine, Chan-Yeung Centre for Occupational and Environmental Respiratory Disease, Vancouver Coastal Health Research Institute, UBC, Vancouver, BC, C

Ilaria Malanchi

The Francis Crick Institute https://orcid.org/0000-0003-4867-3311

Allan Hackshaw

University College London https://orcid.org/0000-0002-5570-5070

Kevin Litchfield

The Francis Crick Institute https://orcid.org/0000-0002-3725-0914

Mariam Jamal-Hanjani

Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute https://orcid.org/0000-0003-1212-1259

James DeGregori

Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, Colorado.

Biological Sciences - Article

Keywords:

Posted Date: September 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1770054/v1

Version of Record: A version of this preprint was published at Nature on April 5th, 2023. See the published version at https://doi.org/10.1038/s41586-023-05874-3.

Abstract

Environmental carcinogenic exposures are major contributors to global disease burden yet how they promote cancer is unclear. Over 70 years ago, the concept of tumour promoting agents driving latent clones to expand was first proposed. In support of this model, recent evidence suggests that human tissue contains a patchwork of mutant clones, some of which harbour oncogenic mutations, and many environmental carcinogens lack a clear mutational signature. We hypothesised that the environmental carcinogen, <2.5μm particulate matter (PM2.5), might promote lung cancer promotion through non-mutagenic mechanisms by acting on pre-existing mutant clones within normal tissues in patients with lung cancer who have never smoked, a disease with a high frequency of EGFR activating mutations. We analysed PM2.5 levels and cancer incidence reported by UK Biobank, Public Health England, Taiwan Chang Gung Memorial Hospital (CGMH) and Korean Samsung Medical Centre (SMC) from a total of 463,679 individuals between 2006-2018. We report associations between PM2.5 levels and the incidence of several cancers, including EGFR mutant lung cancer. We find that pollution on a background of EGFR mutant lung epithelium promotes a progenitor-like cell state and demonstrate that PM accelerates lung cancer progression in EGFR and Kras mutant mouse lung cancer models. Through parallel exposure studies in mouse and human participants, we find evidence that inflammatory mediators, such as interleukin-1 , may act upon EGFR mutant clones to drive expansion of progenitor cells. Ultradeep mutational profiling of histologically normal lung tissue from 247 individuals across 3 clinical cohorts revealed oncogenic EGFR and KRAS driver mutations in 18% and 33% of normal tissue samples, respectively. These results support a tumour-promoting role for PM acting on latent mutant clones in normal lung tissue and add to evidence providing an urgent mandate to address air pollution in urban areas.

Introduction

Barrier organs such as the lung are directly impacted by exposure to environmental challenges. Accordingly, more than 20 environmental and occupational agents are proven lung carcinogens (IARC, 2015). Risk factors driving lung cancer are of particular concern for people who have never smoked. Not only is lung cancer in never smokers (LCINS) the 8th most common cause of cancer death in the UK [1], but these patients are not captured by current screening protocols and no risk-stratification approaches exist for population-based screening. LCINS has distinct clinical and molecular characteristics compared to lung cancer in smokers [2]. In particular, LCINS frequently harbour EGFR oncogenic mutations and are characterised by a significantly lower mutation burden with no clear environmental mutational signature [3–6]. Clonal driver mutations in *EGFR* are more frequent in female patients, and in East Asian compared to Western patients [7]. Several hypotheses have been proposed to explain the observed sex and geographical disparities of *EGFR* mutant lung cancer, including germline genetics [8], ethnicity, radon exposure, occupational carcinogen exposures and air pollution [9].

Ambient air pollution stands out amongst these carcinogens as an estimated 99% of people live in areas that exceed WHO guidelines of 5 μg/m^3. Whilst air pollution levels vary widely between countries, it is the world's fourth leading cause of death, responsible for 6.7 million deaths in 2019 [10]. 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 COPD and asthma [11]. PM is categorized by size including coarse particles with an aerodynamic-mass median diameter, <10 μm, PM_{10}), fine particles <2.5 μm, ($PM_{2.5}$) and ultrafine particles (<0.1 μm, $PM_{0.1}$). $PM_{2.5}$ has been classified as a category 1 carcinogen by the International Agency for Research on Cancer (IARC) and has become increasingly implicated in lung cancer risk [12].

There are established associations between air pollution and overall risk of lip, oral, pharyngeal [13,14], and lung [12,14,15] cancers, including lung cancers in never smokers [16]. A modest increase in *EGFR* mutation rates [17] and an elevated frequency of specific *EGFR* driver mutations (S768I, G719X) [18] have been noted in lung tumours from the severely polluted Xuan-wei region of China. Controlled human exposure studies have found acute diesel exhaust exposure can promote airway inflammation [19]. Macrophages and lung epithelia are the predominant cells that process inhaled PM and cooperatively produce proinflammatory mediators [20]. However, the mechanisms by which PM promotes lung cancer initiation are poorly understood.

Traditionally, it is thought that carcinogens act via mutagenic mechanisms, directly inducing DNA damage [21–23]. However, recent data suggests that not all carcinogens cause a currently detectable mutational signature [24,25]. A recent genetic analysis found that mutational signatures do not fully explain the varied geographical incidence of oesophageal cancer [26], and efforts that have profiled LCINS tumour genomes failed to detect a dominant carcinogenic signal of mutations deriving from exogenous sources [6,27–30]. In particular, the Sherlock study [30] identified exogeneous mutational signatures in only 3% of 232 LCINS genomes. An additional hypothesis for how environmental agents may act is by promoting cancer development from initiated but dormant mutant cells [31]. In the absence of exposure to a promoting agent, mutant cells remain dormant for most of the lifespan of the mouse [32]. In support of this, sensitive deep sequencing approaches have revealed mutations in clones within histologically normal tissues from a range of anatomical sites, a minority of which are known to be driver oncogenic mutations in tumours [33–36].

We hypothesised that air pollution may promote inflammatory changes in the normal tissue microenvironment that might permit mutated nascent clones to expand and initiate tumours. To address

this, we combined epidemiological evidence with functional pre-clinical mouse cancer models, as well as mouse and human $PM_{2.5}$ exposure studies, to decipher potential mechanisms of air pollution-induced lung tumour promotion and actionable targets for cancer prevention (Figure 1A).

Results

Cancer incidence is associated with $PM_{2.5}$ exposure in prospective cohort study analysis

To explore the relationship between air pollution and cancer risk we performed Cox regression analysis on 447,932 UK Biobank participants, where cancer incidence and residential outdoor $PM_{2.5}$ information was available for the year 2010. This demonstrated that $PM_{2.5}$ levels (calculated at 1 $\mu g/m^3$ increments) were associated with lung cancer incidence (HR: 1.16, p<0.001), consistent with a prior report from Huang et al [15]. Our analysis went on to reveal additional associations of $PM_{2.5}$ exposure with the incidence of mesothelioma (HR: 1.19, p=0.032), glioblastoma (HR: 1.19, p=0.015), larynx (HR: 1.26, p=0.020), lip, oral cavity and pharynx (HR: 1.15, p=0.007), small intestine (HR: 1.30, p=0.001), as well as anus and anal canal cancers (HR: 1.23, p=0.031) (Figure 1B). Furthermore, interaction tests between smoking and $PM_{2.5}$ exposure suggest that smoking and high $PM_{2.5}$ levels may have a combined effect on elevating the risk of lung cancer (Supplementary Table S1).

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

While there is a clear tobacco-associated mutational signature in lung cancer in smokers, LCINS is characterized as harbouring relatively few mutations [30] and no clear mutational signature from environmental causes, suggesting alternative mechanisms of LCINS initiation. Thus, we focused our analysis on LCINS, using EGFR mutant cancer as a surrogate of never-smoker status, due to its high prevalence in LCINS [2]. 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 (92.03% Caucasian cohort; $PM_{2.5}$ IQR: 9.95-11.2 µg/m3), South Korea (predominantly Asian cohort: $PM_{2.5}$ IQR: 24.0-27.0 µg/m3) and Taiwan (predominantly Asian cohort; $PM_{2.5}$ IQR: 24.3-38.2 µg/m3). In each country, we consistently observed positive correlations between $PM_{2.5}$ levels (average concentration per geographical area) and estimated EGFR mutant lung cancer incidence in that geographical area: England: R=0.58 p=0.0077 (weighted by the number of cases tested for EGFR

mutations : R=0.32 p=0.17; Figure 1C); Korea: R=0.42 p=0.016 (weighted: R=0.19 p=0.30; Figure 1D); Taiwan R=0.62 p=7.5e-08 (weighted: R=0.37, p=0.003; Figure 1E). Taken together, these epidemiological data, combined with published evidence demonstrating the association between $PM_{2.5}$ and never smoker lung cancer [16], support an association between the estimated incidence of EGFR mutant lung cancer and levels of $PM_{2.5}$.

Air pollution promotes EGFR mutant lung cancer progression in mouse models

Next, we examined whether PM exposure upon mouse models of lung adenocarcinoma could promote tumour development. We induced lung-specific expression of oncogenic human *EGFR*^{L858R} mutations in lung epithelial tissue using a mouse engineered with *Rosa26*$^{LSL\text{-}tTa/LSL\text{-}tdTomato}$*; TetO-EGFR*^{L858R} (ET mice). Upon intra-tracheal delivery of adenoviral Cre recombinase, rare, sporadic lung epithelial cells express oncogenic *EGFR* and expand to develop pre-invasive lesions by 10 weeks. To model exposure of PM, we used PM from the National Institute of Standards and Technologies (NIST), with certified mass fraction values of both organic and inorganic constituents from multiple analytical techniques which represents fine PM from a modern urban environment (Schantz et al., 2016) and administered physiologically relevant doses [37]. Mice were given intratracheal administration of PM or PBS control three times per week for three weeks after the induction of *EGFR*^{L858R}, followed by lesion analysis at 10 weeks post *EGFR*^{L858R} induction (Figure 2A). Analysis of mice at 10 weeks revealed a significant, dose-dependent increase in the number of EGFR mutant cells that had undergone clonal expansions to form early neoplastic lesions in mice exposed to PM (control vs 5 µg p=0. 047; control vs 50 µg p=0.0007; Figure 2B).

In order to ensure that the impact of PM on tumour initiation was not confounded by the mode of *EGFR*^{L858R} induction through adenoviral-cre delivery and to explore the effects of PM on adenocarcinoma promotion, we used the *CCSP-rtTa; TetO-EGFR*^{L858R} model of lung adenocarcinoma. Here, *EGFR*^{L858R} expression is induced in the majority of lung epithelial cells by doxycycline diet and mice develop lung adenocarcinomas within 8 weeks [38]. In this model, at 10 weeks post-induction, we found that mice exposed to 50 µg PM had significantly more lung adenocarcinoma lesions (p=0.032; Figure 2C).

Given the epidemiological data supporting a combined interaction of smoking exposure and $PM_{2.5}$ in the risk of lung cancer, we addressed whether the interactions between PM and lung tumour initiation could also be observed when the initiating oncogene is KRAS, found more commonly in ever-smokers. We used the *Rosa26*$^{LSL\text{-}tdTomato/+}$;*Kras*$^{LSL\text{-}G12D/+}$ (KT) mouse model of lung cancer which generates tumours upon intratracheal viral delivery of Cre recombinase that rarely progress beyond adenoma and accurately mimics human lung adenoma at molecular and histopathological levels [39]. Administration of both 5 µg

and 50 µg PM significantly increased the number of early neoplastic lesions in this model at 10 weeks post induction of oncogene compared to control (5 µg p=0.048; 50 µg p=0.0087;Figure 2D). These data suggest that PM can promote the expansion of EGFR mutant cells to pre-invasive lesions and the formation of lung adenocarcinoma, and that these effects are observed for both oncogenic *Kras* and *EGFR*-driven lesions.

To begin exploring the mechanisms by which PM promotes EGFR mutant lung tumourigenesis, we tested 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 [40] 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 2A), 3 weeks of exposure to PM did not result in a significant increase in neoplastic lesions, suggesting a competent immune system is required for PM-enhanced EGFR mutant lung tumourigenesis (p=0.879; Figure 2E).

The inhalation of toxic particles induces a local response in the lung which is mediated by macrophages and lung epithelial cells (Hiraiwa & van Eeden;Hogg & Van Eeden), we therefore profiled both the myeloid and epithelial response to PM in immune competent lungs harbouring EGFR mutant cells (ET mice) or not (T mice, *Rosa26*$^{LSL\text{-}tdTomato/+}$) . We exposed adenoviral-Cre recombined T and ET mice to 3 weeks of PM in vivo, harvesting animals 24 hours after the final exposure to analyse the acute immune response. We observed a marked increase in the proportion of interstitial macrophages (IMs)(T p=0.0427, ET p=0.0335; Figure 2F) and the expression of PD-L1 upon these cells in both T and ET mice (T p=0.0309, ET p=0.0061; Figure 2G). There was no difference in alveolar macrophage proportion in the lung but a significant increase in neutrophils in T mice only, whereas dendritic cells were only elevated in ET mice (Supplementary Figure S1A). To extend the flow cytometry findings we carried out immunofluorescence staining of ET lungs using the pan-macrophage marker CD68. We observed increased density of CD68+ macrophages with PM exposure both at early and later timepoints, suggesting retention of PM-associated IMs throughout early tumourigenesis (3 weeks p=<0.0001; 10 weeks p=0.0217; Figure 2H,I). These data support the hypothesis that PM exposure is associated with enhanced macrophage infiltration in the lung, in line with previous results in humans [20].

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

Next, to understand how PM may affect both healthy lung epithelium and epithelium harbouring EGFR mutations, we carried out RNA-seq of sorted and 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). We observed that particulate matter induced significant alterations in the transcriptome of epithelia from both T and ET mice, with PM accounting for 19% of the variance in differentially expressed genes and EGFR mutation accounting for 38% of the variance (Figure 3A). 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 (Figure 3B; Supplementary Figure S1B). In particular, we observed upregulation of inflammatory cytokines, including interleukin-1β (IL1β), GM-CSF (CSF2) and IL33 (Figure 3C). Lung injury models in mice can induce cell state changes within alveolar type II (AT2) cells, a likely cell of origin of lung adenocarcinoma [41], and expand populations with a progenitor-like phenotype which mediate alveolar regeneration [42,43]. Consistent with this, we noted upregulation of genes previously associated with altered, progenitor-like AT2 cell states (Figure 3C).

To understand the relevance of these transcriptional changes to human lung, we explored RNA-seq data from a clinical crossover study in which lung brushings, containing lung epithelial cells, were collected from 9 individuals who have never-smoked [44]. Samples were taken after exposure to diesel exhaust for 2 hours and exposure to filtered air control conditions with a 4 week washout period between exposures in the same individual [45] (Figure 3D). Diesel exhaust is a robust model of traffic-related air pollution and a dose approximating 300 μg/m^3 of $PM_{2.5}$ was used to represent exposure levels documented in polluted mega-cities and occupational exposures [46]. Using a custom geneset established by selecting human orthologs of genes significantly differentially expressed between T vs T-PM condition in our mouse model, we compared the fold change of gene expression with pollution exposure in the T mouse lungs to the fold change of gene expression with diesel exposure in never-smokers (Figure 3E). We found significant upregulation of the inflammatory marker *Lipocalin-2* (*LCN2)* and anti-inflammatory *SLPI* which inhibits neutrophil elastase [47]. We also noted that immune-related genes (*IL4I1*, *CXCL3*, and *IL1β*) and genes associated with a perturbed AT2 cell state (*ORM1*, *ITGA7* and *LRG1*) showed similar trends and directionality in the human exposure study, although did not reach statistical significance. These results suggest that there may be similar gene expression programs induced following PM exposure between species, converging on certain inflammatory mediators and upregulation of genes associated with an AT2 cell progenitor-like state.

These results identify PM induced inflammatory pathways in mice and humans and transcriptional changes associated with lung progenitor cell states [42]. To test if these transcriptional changes associated with progenitor-like states are reflected in functional differences in epithelial cell behaviour following PM exposure, we performed a lung organoid formation assay [48] in which lung epithelial cells from ET mice were isolated and grown as 3D organoids *ex vivo* following *in vivo* exposure to PM (Figure 3F). Whilst

there was a trend for non-recombined cells within ET mice exposed to PM to have increased organoid formation efficiency (OFE)(p=0.0747; Figure 3G,H); recombined, tdTomato+ *EGFR*^{L858R} cells demonstrated a more pronounced and significant increase in OFE (p=0.0245; Figure 3G,H). This suggests that the combination of PM and cells harbouring the *EGFR*^{L858R} driver mutation increases progenitor function of oncogenic epithelial cells that is not seen with PM exposure or expression of mutant *EGFR* alone.

Lung epithelial cells and macrophages generate a complex milieu of inflammatory mediators when exposed to particulate matter [20]. One key mediator is IL1β, which was upregulated in PM exposed epithelia by RNA-seq. IMs are a major source of IL1β and we observed an increased number and activation of IMs in T and ET mice exposed to PM [42]. Therefore, we explored whether treatment with IL1β is sufficient to promote expansion of EGFR mutant organoids. AT2 cells were isolated from ET mice not exposed to PM, followed by oncogene activation *in vitro* with adenoviral-Cre incubation, and plated in the organoid assay with *in vitro* IL1β treatment. This resulted in an expansion of organoid size (p=0.0012), with organoids expressing markers of both differentiated AT2 cells (SPC) and progenitor-like AT2 cells (Keratin 8) respectively (Figure 3I, J). These data suggest that IL1β is capable of expanding *EGFR* mutant epithelial cells with progenitor-like capacity.

EGFR and *KRAS* mutations are frequently found in histologically normal lung tissue

If tumour development does occur via two stages, initiation and promotion, this is contingent on the cell harbouring a pre-existing oncogenic driver mutation [31]. 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 in the skin, oesophagus, bladder and liver. (Supplementary Table S2). Therefore, we sought evidence for *EGFR* driver mutations in normal lung tissue distinct from those present in matched lung tumour in people with lung cancer, other cancers and no evidence of cancer, using digital droplet PCR (ddPCR) or Duplex-seq (Figure 4A, Supplementary Table S3).

Firstly, we selected normal lung tissue from 195 of 1346 prospectively recruited TRACERx (NCT01888601) patients, 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 S3, Supplementary Figure S3). 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% [49]

We filtered out instances that had 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 suggest that *EGFR* driver mutations can be present in normal lung tissue, even in patients where the same mutations were not selected for during tumourigenesis.

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, can act as a surrogate for exposure to ambient air pollution [50]. 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). Whilst there are multiple environmental contributors to anthracosis, including smoking and inhaled pollutants, these data suggest pollutants do not enhance the frequency of activating oncogenic mutations but rather promote the expansion of pre-existing clones.

Next, we addressed 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 the time of autopsy within the PEACE (NCT03004755) study from 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 S3, Supplementary Figure S3). An *EGFR* mutation was detected in the normal lung of 3 out of 19 (16%) patients (Figure 4B). Despite spatially separated multi-region profiling of normal tissue in 15 of the 19 patients, mutations were detected in only 1 region in these 19 patients.

We attempted to validate this finding using an independent ultra-deep sequencing platform in additional cohorts of patients with and without lung cancer, addressing whether driver mutations existed at other genomic loci in *EGFR* and in *KRAS*. Using Duplex-seq, we analysed 33 normal lung tissue samples derived from the Biomarkers and Dysplastic Respiratory Epithelium (BDRE) Study (NCT00900419, Figure 4A, Supplementary Table S4, Supplementary Figure S3). 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 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, and *KRAS* GTP binding domain exons 2 and 3, with a limit of detection of <0.01%. Given the broader range of *EGFR* mutations detected by Duplex-seq across several exons, we only considered mutations featured in the cancer gene census [51], and further filtered mutations by evidence of driver mutation status in the literature (Supplementary Table S2). In 15 of 20 cancer cases where tissue was available, we also performed Duplex-seq on the suspicious involved nodule/cancer identified in the CT scan, to confirm that the mutations present in normal tissue were found exclusively in the normal lung tissue samples. 11/33 (33%) samples harboured a *KRAS* driver mutation (G12X, G13X, Q61X; Figure 4E), while 4/33 (12%) samples harboured an *EGFR* driver mutation (E709K, G719D, T725M; Figure 4F).

In summary, Duplex-seq and ddPCR revealed that 45/247 (18%) of normal lung samples harboured an *EGFR* driver mutation, and 11/33 (33%) normal lung samples harboured a *KRAS* driver mutation. When we compared proportions of samples that harboured *EGFR* or *KRAS* mutations, we did not see any obvious trends between sex, smoking, or diagnoses groups. (Figure 4G) suggesting mutations accumulate in normal lung independently of these clinical characteristics.

Discussion

70 years ago, Berenblum and Shubik developed the concept of two processes involved in carcinogenesis; tumour initiation, whereby cells acquire oncogenic driver mutations and tumour promotion, 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 [31]. This suggests cancer development is driven by cells harbouring oncogenic mutations in histologically normal tissues and environmental/inflammatory stimuli driving tumour promotion and overt malignancy [32]. A number of risk factors have been identified for LCINS including second-hand smoke, occupational carcinogen exposure, germline genetics [8] and radon exposure [9]. In this study, we explored the paradigm of tumour promotion in the development of lung cancer by air pollutants. LCINS tumours typically have a low mutational burden and no discernible, common exogenous mutational source despite 99% of people living in areas that exceed WHO guidelines of 5 $\mu g/m^3$ generating the hypothesis that PM may act as a tumour promoter in this context [3,4,30].

We find that PM is associated with an increased risk of *EGFR* mutant lung cancer. *EGFR* mutant lung cancer incidence and $PM_{2.5}$ are elevated in East Asian countries, consistent with the biased geographical distribution of LCINS in Asia [52]. A limitation of our analysis is its ecological nature: the analysis was performed at the geographical region level as patient-level data for *EGFR* mutation status were not available for all cohorts. In addition, $PM_{2.5}$ levels were only summarised at a geographical region level, as these were the only data available for all three within-country cohorts.

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 the body of research identifying mutations within a range of histologically normal tissues [33–36]. These *EGFR* and *KRAS* mutations are found at similar frequencies in normal lung tissue from patients with an established diagnosis of lung cancer (TRACERx Study Cohort) and from patients who do not acquire lung cancer in their lifetime (PEACE Study Cohort). We observed that PM fosters an AT2 cell state with progenitor function in *EGFR* mutant cells from mice. 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 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. This is in agreement with the observation of the rapid progression of mutant cells to carcinoma in mice following treatment with a promoter agent even after 1 year of acquiring the mutation [32].

Our results provide additional evidence that a major risk factor for cancer development is not only the almost inevitable acquisition driver mutations in normal epithelium but also mechanisms (both intrinsic and extrinsic) that promote nascent mutant cell expansion. 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 causes of environmental carcinogenesis. A broader approach will be necessary to establish potential hormonal, environmental and germline influences that might promote or restrict mutant clone expansions. This will involve the collection of global environmental exposure datasets, longitudinal cohorts studying at-risk populations and human and pre-clinical studies to understand how exposures perturb normal tissue physiology and permit mutant clone expansions. 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 “promotion” 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 [53].

In the short term, these data suggest a mechanistic and causative link between pollution and lung cancer, first proposed by Doll and Hill in 1950 [54], providing a public health mandate to urgently restrict particulate emissions in urban areas.

Declarations

Acknowledgements

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. 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. 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). The results published here are based in part on data generated by The Cancer Genome Atlas pilot project established by the NCI and the National Human Genome Research Institute. 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 and the Biological Resources Unit. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.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).

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 analyses. C.L. performed the human RNA-seq analyses and curated the pollution data. C.W. performed the mouse experiments and curated the mutation literature. 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. H.C. and S.-H.L. performed tne Korea epidemiological analyses. F.V.M, J.B., M.A. and D.C. were involved in mouse data acquisition. 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. 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, M.A.B. and C.B. wrote and ran the >i-seq 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. N.M. provided supervision over bioinformatics analyses. C.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.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).

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).

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. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).

4. Devarakonda, S. *et al.* Genomic Profiling of Lung Adenocarcinoma in Never-Smokers. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **39**, 3747–3758 (2021).

5. Govindan, R. *et al.* Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. *Cell* **150**, 1121–1134 (2012).

6. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **376**, 2109–2121 (2017).

7. 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).

8. 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).

9. Couraud, S., Zalcman, G., Milleron, B., Morin, F. & Souquet, P.-J. Lung cancer in never smokers–a review. *Eur J Cancer* **48**, 1299–1311 (2012).

10. 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* **397**, 2337–2360 (2021).

11. 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* **389**, 1907–1918 (2017).

12. Ciabattini, M., Rizzello, E., Lucaroni, F., Palombi, L. & Boffetta, P. Systematic review and meta-analysis of recent high-quality studies on exposure to particulate matter and risk of lung cancer. *Environ Res* **196**, 110440 (2021).

13. Chu, Y.-H. *et al.* Association between fine particulate matter and oral cancer among Taiwanese men. *J Investig Med* **67**, 34–38 (2019).

14. Coleman, N. C. *et al.* Fine Particulate Matter Exposure and Cancer Incidence: Analysis of SEER Cancer Registry Data from 1992-2016. *Environ Health Perspect* **128**, 107004 (2020).

15. 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).

16. Myers, R. *et al.* High Ambient Air Pollution Exposure Among Never Smokers Versus Ever Smokers with Lung Cancer. *J Thorac Oncol* S1556-0864(21)02256–5 (2021) doi:10.1016/j.jtho.2021.06.015.

17. Yu, X.-J. *et al.* Characterization of Somatic Mutations in Air Pollution-Related Lung Cancer. *EBioMedicine* **2**, 583–590 (2015).

18. Lv, L. *et al.* Distinct EGFR Mutation Pattern in Patients With Non-Small Cell Lung Cancer in Xuanwei Region of China: A Systematic Review and Meta-Analysis. *Front Oncol* **10**, 519073 (2020).

19. 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).

20. 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).

21. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

22. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

23. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat Commun* **9**, 1744 (2018).

24. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16 (2019).

25. Riva, L. *et al.* The mutational signature profile of known and suspected human carcinogens in mice. *Nat Genet* **52**, 1189–1197 (2020).

26. Moody, S. *et al.* Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. *Nat Genet* **53**, 1553–1563 (2021).

27. 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).

28. Lee, J. J.-K. *et al.* Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma. *Cell* **177**, 1842-1857.e21 (2019).

29. Wang, C. *et al.* Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. *Nat Commun* **9**, 2054 (2018).

30. Zhang, T. *et al.* Genomic and evolutionary classification of lung cancer in never smokers. *Nat Genet* **53**, 1348–1359 (2021).

31. 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).

32. Balmain, A. The critical roles of somatic mutations and environmental tumor-promoting agents in cancer risk. *Nat Genet* **52**, 1139–1143 (2020).

33. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).

34. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).

35. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).

36. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).

37. Chan, Y. L. *et al.* Pulmonary inflammation induced by low-dose particulate matter exposure in mice. *Am J Physiol Lung Cell Mol Physiol* **317**, L424–L430 (2019).

38. 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).

39. 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).

40. McDaniel Mims, B. & Grisham, M. B. Humanizing the mouse immune system to study splanchnic organ inflammation. *The Journal of Physiology* **596**, 3915–3927 (2018).

41. Sutherland, K. D. *et al.* Multiple cells-of-origin of mutant K-Ras–induced mouse lung adenocarcinoma. *Proceedings of the National Academy of Sciences* **111**, 4952–4957 (2014).

42. 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).

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

44. 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.

45. 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).

46. Wooding, D. J. *et al.* Particle Depletion Does Not Remediate Acute Effects of Traffic-related Air Pollution and Allergen. A Randomized, Double-Blind Crossover Study. *Am J Respir Crit Care Med* **200**, 565–574 (2019).

47. Camper, N. *et al.* A secretory leukocyte protease inhibitor variant with improved activity against lung infection. *Mucosal Immunol* **9**, 669–676 (2016).

48. 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).

49. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* **3**, lqab019 (2021).

50. 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).

51. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941–D947 (2019).

52. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).

53. 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).

54. Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br Med J* **2**, 739–748 (1950).

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 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.

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)

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

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) Duplex-seq of samples from the BDRE study

Normal lung tissue samples

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 Duplex-seq. The absence of nodules or masses detected by chest CT scans was indicative of the non-tumor nature of these contralateral samples. To document that the brushings were peripheral, they were 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 Duplex-Seq)

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), 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.

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. Duplex-seq reads were processed using an in-house pipeline adapted from Valentine et al.[5] Additionally, we also profiled the involved lung of 15 of 20 cases where the suspicious nodule in the contralateral lung was cancerous, and where tissue was available. These data were processed by the bioinformatics pipeline provided by TwinStrand BioSciences. Using these, 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 Duplex-seq 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 Duplex-seq 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

The UK Biobank study comprises over 500,000 participants, aged between 40-69 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].

Following a similar method to that described in[7], we first excluded all participants who had missing particulate matter or genetic principal components data. Multiple imputation with chained equations[8] was used to impute missing values for the remaining 447,932 participants. The imputation model used the following variables: PM2.5, PM2.5-10, PM10, sex, BMI, ever smoking status, passive smoking (weekly hours of tobacco exposure at home), household income (dichotomised into “below” or “greater than or equal to” £31,000 annually), educational attainment (split into “below” or “degree level and above”), and the first 15 genetic principal components (to account for ethnicity). We imputed the dataset using predictive mean matching and logistic regression for continuous and binary variables, respectively, performing a maximum of 90 iterations. This yielded 5 complete versions of the original dataset in which the missing values have been imputed. Convergence was assessed through inspecting the resulting plot. Each imputed dataset was independently used in the same analysis protocol.

Participants were followed up from recruitment until either date of each cancer diagnosis or censoring, which was defined as the time of death or latest date of cancer diagnosis, whichever was earlier. We created a multivariate Cox regression model for each imputed dataset and primary cancer type with >= 100 cases, and pooled results across these models, which were consistent for each cancer type, into a single set using Rubin’s rules[8]. These models included the same covariates as in the imputation model, with the addition of age at the end of follow-up for each cancer. For cancers of the larynx or lip, oral cavity and pharynx, we further corrected for alcohol consumption, excluding those participants with missing alcohol data due to the high missingness of these variables. Schoenfeld residuals were examined to assess the proportional hazards assumption and variables that failed to satisfy this assumption were modelled as time-dependent. Cancer types for which this could not reliably be

performed were excluded. Individual models that failed to converge were not included, and if all models for a particular cancer type failed, then that cancer type was excluded. In total, we thus excluded uterine, acute myeloid leukaemia, melanoma, and non-melanoma skin cancers, as well as 4 models from CRC, 3 from renal (excluding pelvis), and 1 from malignant immunoproliferative disease.

An interaction test between PM2.5 and smoking was performed for lung cancer. The approach described above was used to create individual multivariate Cox regression models for each imputed dataset and aggregate the results.

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 PM2.5 air pollution data (μg/m^3) from 2008 to 2017 was obtained at the grid code level (1km x 1km) from DEFRA[9]. 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, PM2.5 concentration estimates were then aggregated to the Cancer Alliance region level and then averaged over the period 2008 to 2017- these were selected because they represented the 10 years prior to a lung cancer diagnosis in 2018. 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 39290 lung cancers (International Classification of Diseases codes C33 to C34) diagnosed in England between 1 January 2018 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: 92.03%, Asian: 1.47%, Chinese: 0.26%, Black: 1.19%, Mixed: 0.29%, Other: 1.10%, Unknown: 3.68%.

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 CAS2107 monthly snapshot]. Only cases with "Overall: TS" as "a:abnormal" and "b:normal" for EGFR were used in the calculation for EGFR mutation rate (n=8585). 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. PM2.5 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 = <# EGFRm>/(<# EGFRm> + <# EGFRwt>)

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]. $PM_{2.5}$ ($\mu g/m^3$) concentration estimates were available for each county in Taiwan from 2006 to 2017. We averaged $PM_{2.5}$ levels across the 5-year period (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. 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-standardlized 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 of more than 5% of the total population, these were the counties that were retained for analysis.

EGFR mutation rate = <# EGFRm>/(<# EGFRm> + <# EGFRwt>)

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

Analyses were performed separately for each of the four 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 Pearson correlation tests.

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 (<10% 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.

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. R26tTA mice were obtained from Jackson laboratory. Mice were backcrossed onto a C57Bl6/J background and further crossed to generate $Rosa26^{LSL\text{-}tTa/LSL\text{-}tdTomato}$/Tet(O)$EGFR^{L858R}$ mice. Rosa26rtTa/TetO-$EGFR^{L858R}$ and LSL-$Kras^{G12D}$ mice have been described previously[14,15] . 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.Recombination was initiated by adenoviral Cre (Viral Vector Core, University of Iowa, USA) delivered via intratracheal intubation (single dose, $2.5x10^7$virus particles/50 µl).

For exposure to fine particulate matter or control, SRM2786 from the National Institute of Standards and Technologies (NIST) resuspended in sterile PBS using sonication and particle size distribution was confirmed using a zetasizer. Mice were briefly anesthetized using 5% isoflurane and intratracheal administration of 5 µg, 50 µg or control PBS was carried out and recovery monitored. 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).

Fluorescence-activated cell sorting analysis and cell sorting

Mouse lungs were cut 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 70 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 antibody 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 and analysed using a BD Symphony flow cytometer. Data was analysed using FlowJo (Tree Star).

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: EGFR L858R mutant specific (Cell Signaling: 3197, 43B2), anti-RFP (Rockland: 600-401-379) and CD68 (ab283654). Primary antibodies were detected using biotinylated secondary antibodies and detected by HRP/DAB or . Slides were imaged using a Leica Zeiss AxioScan.Z1 slide scanner.

RNA-Sequencing (RNA-seq)

Lung CD45−CD31−Ter119−EpCAM+ were sorted from control and PM exposed mice after PM exposure 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 (Illumina), to achieve an average of 25 million reads per sample.

RNA-seq Analysis

The RNA-seq pipeline of nf-core framework version 3.3 was launched with Nextflow version 21.04.0 to analyse RNA sequencing data[16]. Raw reads in fastq files were mapped to GRCm38 with associated ensemble transcript definitions using STAR version 2.7.6a[17]. 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)[18]. Differential expression analysis was performed using the R platform version 4.0.3 package LIMMA version 3.44.1 filtering with the absolute value of log fold change more 1 and p-value less than 0.05[19]. The gene expression between treatment groups was further analysed for their pathway enrichments using Gene Set Enrichment Analysis (GSEA).

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

RNA sequencing was applied to 18 samples of lung brushings from 9 never-smokers from the COPA study after exposure to filtered air and diesel exhaust. Salmon[20] was used to estimate transcript-level abundance from RNA-seq read data. Differential expression analysis was performed using DESeq2[21]. The log two fold change in gene expression before and after exposure to filtered air and diesel exhaust was calculated. P-values were adjusted using the Benjamini-Hochberg method. The log two fold change of significantly differentially expressed genes from the T control mouse was compared to the log two fold change expression of the genes from COPA participants.

Organoids

Lung tissue was minced manually with scissors and 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 and centrifuged at 1,250 r.p.m. for 10 min. The cell-pellet was incubated in Red Blood Cell Lysis buffer (Miltenyi Biotec) for 5 min at room temperature and passed through a 40 µm filter. After centrifugation, cells were washed with magnetic-activated cell sorting (MACS) buffer (0.5% BSA and 250 mM EDTA in PBS) and passed through a 20 µm strainer-capped tube to generate a single-cell suspension. Antibody staining was then performed for cell isolation or for flow cytometry analysis.

Lung organoid co-culture assays have been previously described[22]. Lung epithelial cells (EpCAM+CD45−CD31−Ter119−) from control or PM exposed mice underwent fluorescence-activated cell sorting (FACS) and were resuspended in 3D organoid media (DMEM/F12 with 10% FBS, 100 U

ml−1 penicillin-streptomycin and insulin/transferrin/selenium (Merck Sigma-Aldrich)). Cells were mixed with murine normal lung fibroblast (MLg) cells and resuspended in GFR Matrigel at a ratio of 1:1. Then 100 µl of this mixture was pipetted into a 24-well transwell insert with a 0.4 µm pore (Corning). In each insert, 2,000-5,000 epithelial cells and 25,000 MLg 2908 cells were seeded. After incubating for 30 min at 37 °C, 500 µl 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 interleukin-1-beta ex vivo treatment of lung alveolar type II cells, digested lung from ET mice (without in vivo Cre induction) was prepared as described above. Alveolar type II cells (AT2) were sort purified as previously described (MHC Class II+CD49f^{low}EpCAM+CD45−CD31−Ter119−)[23] and 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[24]. 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[25] 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.

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.

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

12. > > | . https://ncc.re.kr/cancerStatsList.ncc?sea.

13. . - . https://airtw.epa.gov.tw/CHT/Query/His_Data.aspx.

14. 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).

15. 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).

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

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

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

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

20. 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).

21. 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).

22. 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).

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

24. 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).

25. 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).

Figures

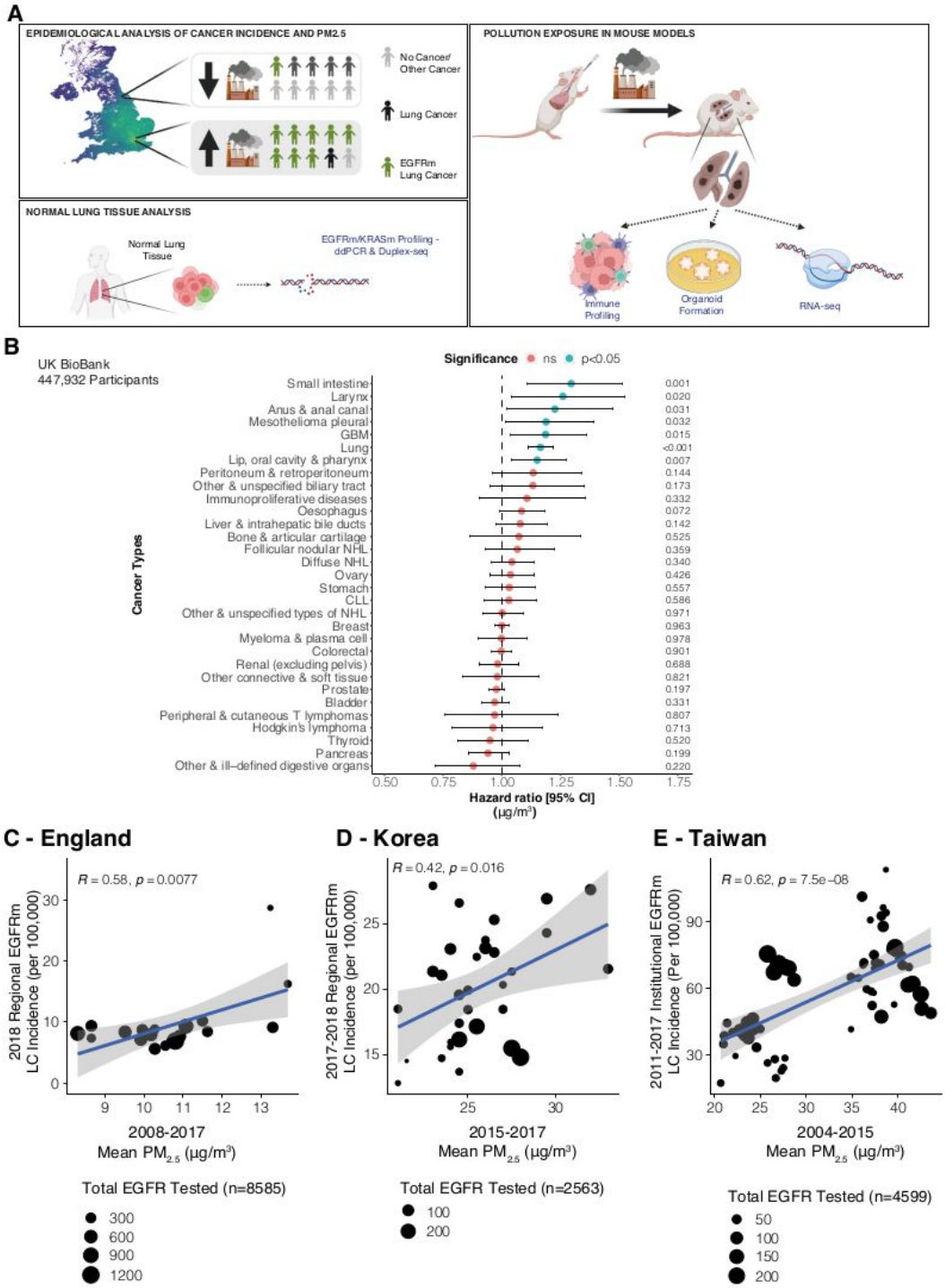

Figure 1

Exploring the association between cancer and air pollution. A) Study design. B) Forest plot indicating the relationship between cancer risk and residential PM2.5 exposure levels (range: 8.17 - 21.31 µg/m3) in the UK Biobank dataset. Cancer types with risk levels that are significantly associated with PM2.5 are indicated with blue dots. HR are reported in units of 1 µg/m3. For colorectal, chronic lymphocytic leukaemia (CLL), as well as anus and anal cancers, the Kaplan-Meier curves may depart from the

proportional hazards assumption at the ends. C-E) Scatter plots showing relationships between PM2.5 and estimated EGFR mutant lung cancer incidence at the country level in England (C), Korea (D) and Taiwan (E).

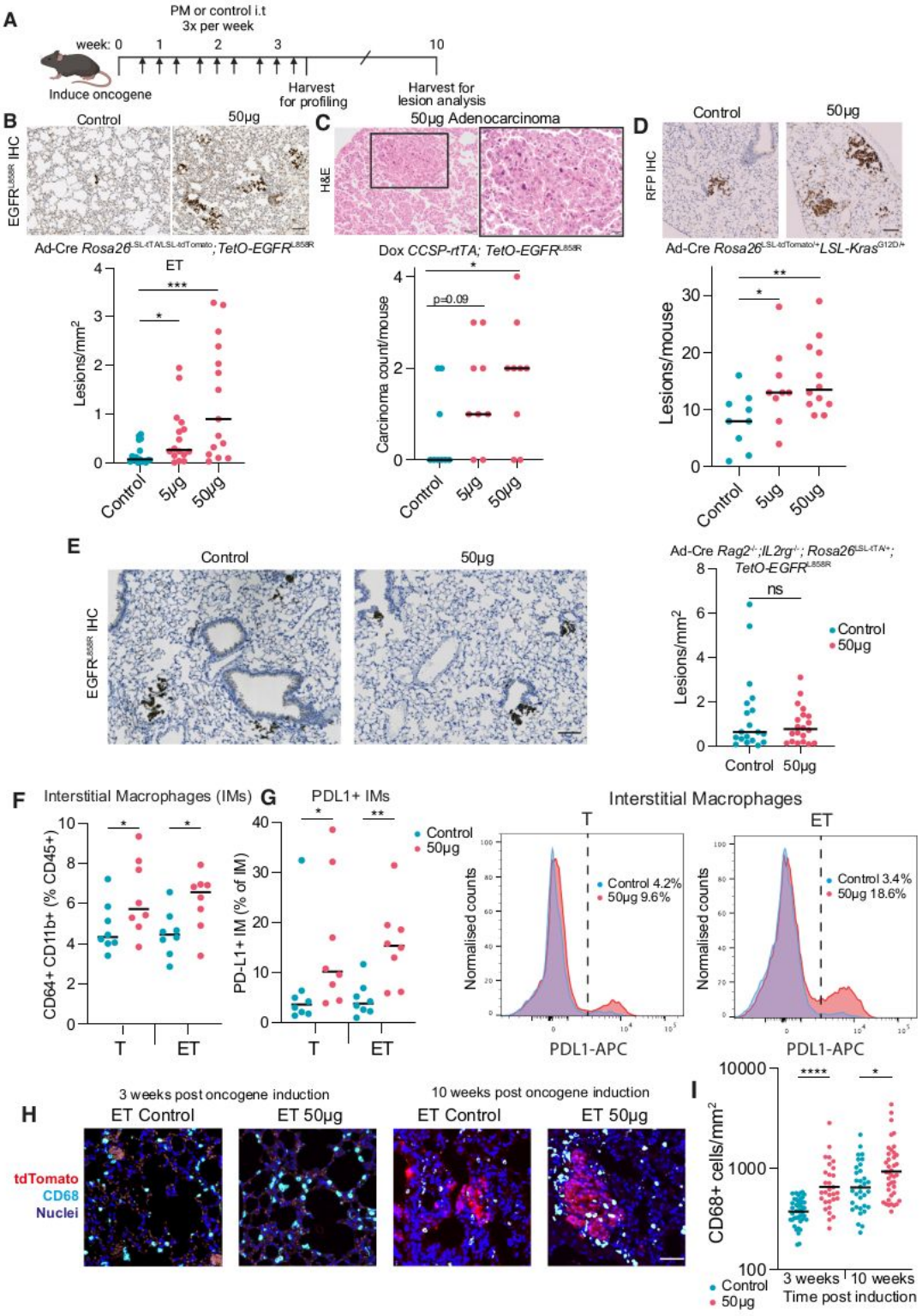

Figure 2

PM promotes lung tumourigenesis. A) Schematic of mouse models of lung cancer indicating induction of oncogene, followed by exposure to particulate matter (PM) and tissue collection for analyses. B) Representative immunohistochemistry (IHC) of human EGFRL858R in control and PM exposed ET mice, with quantification of huEGFRL858R+ pre-cancerous lesions/mm2 of lung tissue below (n=16 control & 5 µg group, n= 15 for 50 µg group). C) Representative H&E of a lung adenocarcinoma in a 50 µg PM exposed, doxycycline treated CCSP-rtTa; TetO-EGFRL858R mice; quantification of number of adenocarcinomas per mouse below (n = 9 per group). D) Representative IHC for red fluorescent protein (RFP, marks tdTomato+ cells) in Rosa26LSL-tdTomato/+;KrasLSL-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). E) Representative IHC of huEGFRL858R in control and PM exposed immune-deficient E mouse models; quantification of number of huEGFRL858R+ pre-cancerous lesions per mm2 of lung tissue (n=19 control and 20 PM exposed) F) Proportion of interstitial macrophages 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). G) Proportion of PD-L1+interstitial macrophages in lung tissue from T and ET mice 24h after final control or PM exposure, representative histogram demonstrating PD-L1 expression within lung interstitial macrophages in T(left) and ET

(right) mice in control (blue) or PM-exposed (pink) conditions. H) 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. I) Quantification of CD68+ cells per mm2 of lung tissue, selecting >30 random fields of view of 500 µm2 (n= 4 mice per group). Gating strategies for flow cytometry analysis provided in Supplementary Figure S2. Statistical analysis by one-way ANOVA for B, C, D, E, F, G & I.

Mann-Whitney for E. Scale bars 100 µm (B,D,E), 50 µm (C main, H), 20 µm (C insert). *p<0.05,

**p<0.01, ***p<0.001.

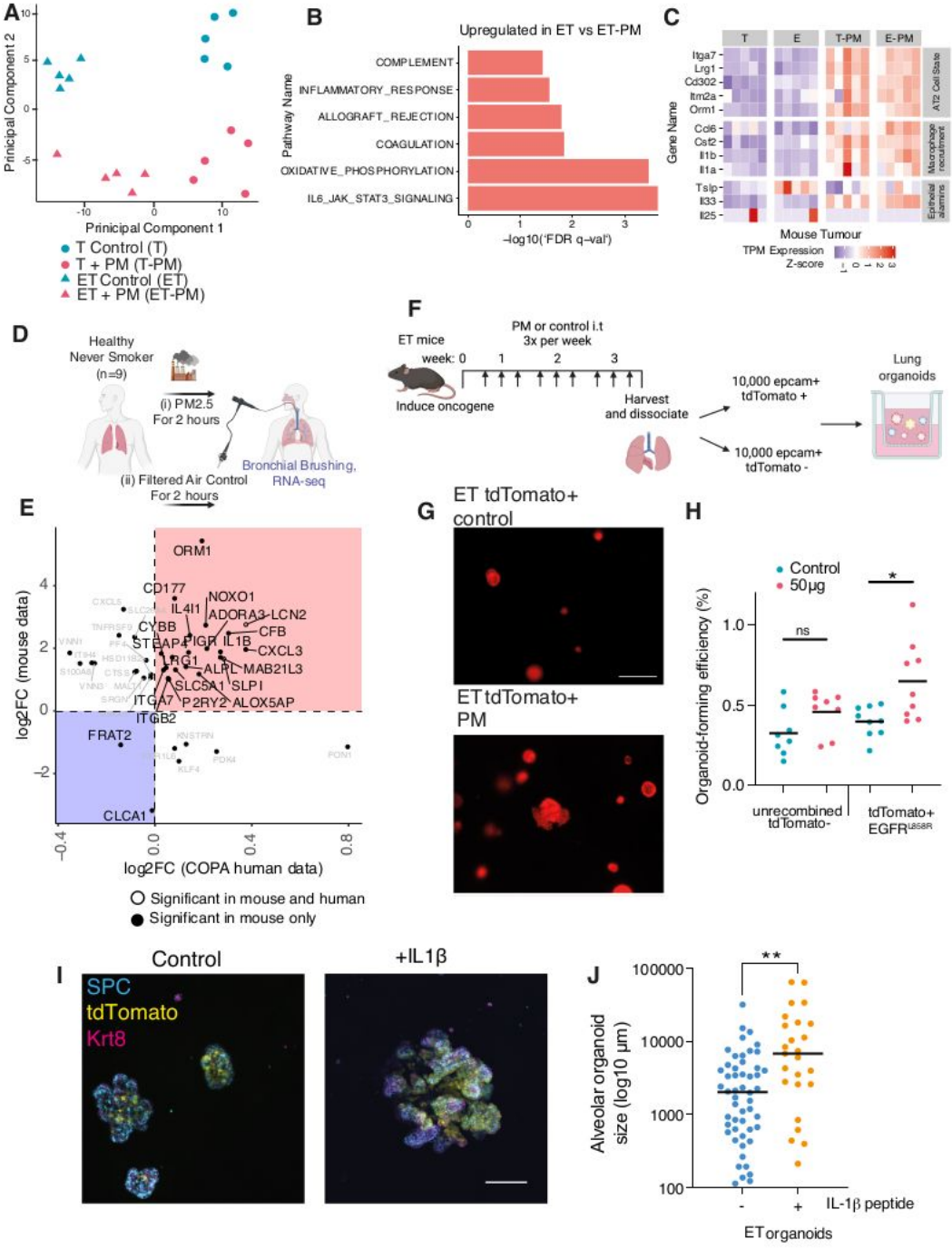

Figure 3

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. D)

Schematic displaying experimental set-up of clinical exposure study in never-smoker volunteers initially reported in (Ryu, 2021), crossover design with (i) and (ii) in random order separated by 4-week washout. E) Fold change (FC) of significantly upregulated genes (identified in mouse) compared to the fold change of genes changed in the clinical exposure study. With common directionality across species indicated (negative: grey background; positive: red background). F) Schematic describing in vivo exposure of ET mice to control or PM, followed by isolation of EpCAM+ tdTom+EGFRL858R+ cells and EpCAM+tdTom-EGFRL858R- cells and plating in epithelial organoid assay. G) Representative fluorescent images of tdTomato organoids at day 14 from control ET mice or ET mice exposed to pollution in vivo. H) 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)). I) 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. I) 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. J. Statistical analysis by one-way ANOVA for H and Mann-Whitney for J. Scale bar 500 μm (G), 50 μm (I). *p<0.05, **p<0.01, ***p<0.001.

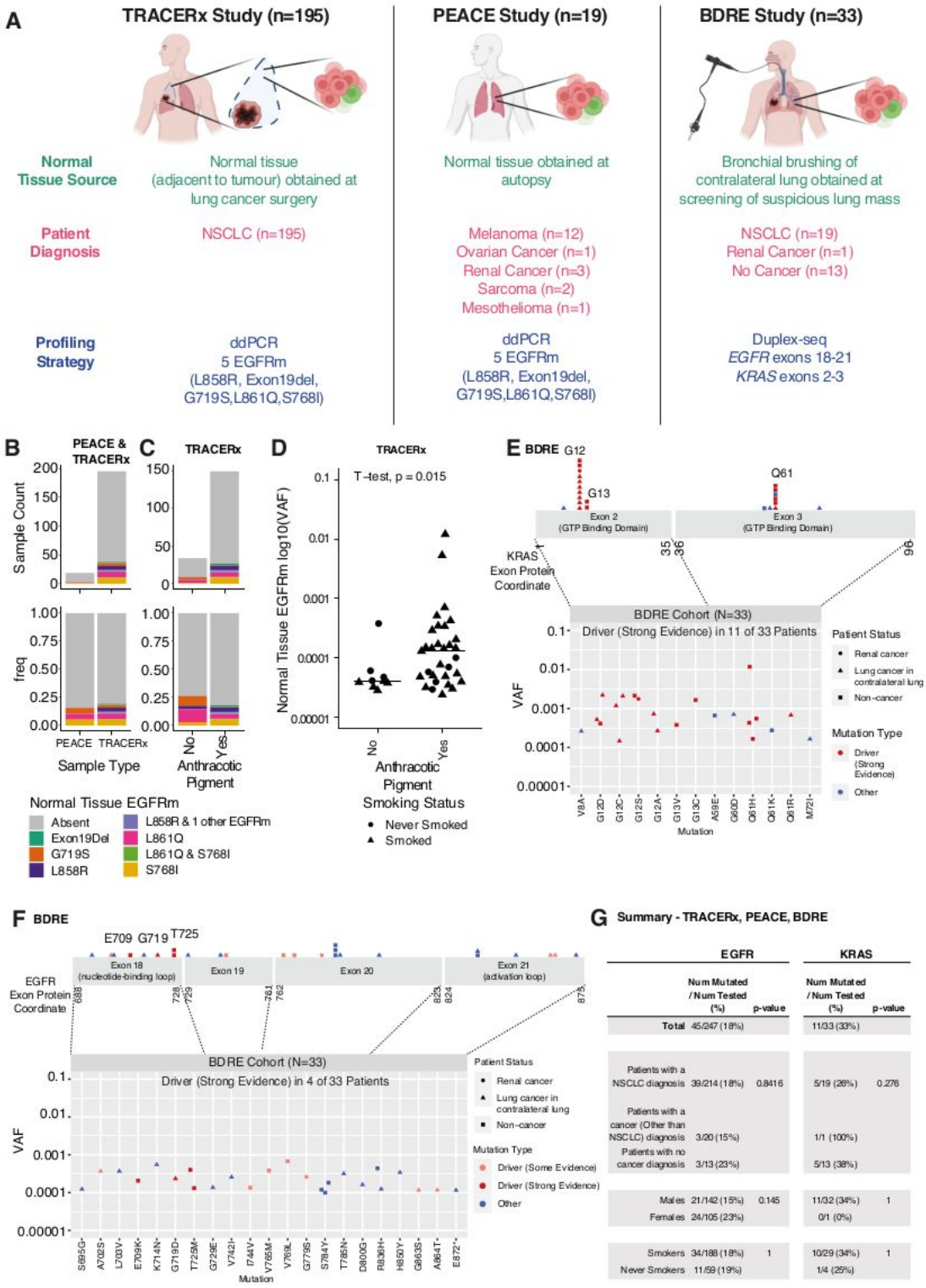

Figure 4

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, separating samples with and without anthracotic pigment. E) 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. F) Top: EGFR Mutations detected using Duplex-seq across KRAS exons 2-3 on normal lung samples from the BDRE Study. Bottom: VAFs of each EGFR mutation are displayed. 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. (Details of driver mutations can be found in Supplementary Table S5) G) Summary table of EGFR and KRAS mutation identified across all three cohorts (TRACERx, PEACE, BDRE).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1UKBiobankInteractionTestsCancerTypeDefinitions.xlsx
- SupplementaryTableS2ExistingEffortsofProfilingofMutationsinNormalTissues.xlsx
- SupplementaryTableS3TRACERxandPEACEStudiesddPCRCohortClinicalCharacteristics.xlsx
- SupplementaryTableS4BDREStudyDuplexseqCohortClinicalCharacteristics.xlsx
- SupplementaryTableS5EvidenceofEGFRDriverMutationStatus.xlsx
- SupplementaryTableS6Reagents.xlsx
- SupplementaryFigures.pdf