nature portfolio

Corresponding author(s): Charles Swanton

Last updated by author(s): 10/02/2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data
Data analysis	R (version 3.6.2) Alignment and QC: FastQC (version 0.11.8) FastQ Screen (version 0.13.0) bwa-mem (version 0.7.17) Sambamba (version 0.7.0) Picard Tools (version 2.21.9) GATK (version 3.8.1) Somalier (version 0.2.7) Samtools (version 1.9) Conpair (version 0.2) Variant Calling: SAMtools (version 1.10) VarScan2 (version 2.4.4) MuTect (version 1.1.7) bam-readcount (version 0.7.4) Annovar (version: Revision 529) R packages used in version 3.6.3: fst (version 0.9.4) tidyverse (version 1.3.0) survival (version 3.2.13)

ggplot2 (version 3.3.2)
dplyr (version 1.0.2)
tidyr (version 1.1.2)
gridExtra (version 2.3)
cowplot (version 1.1.0)
survminer (version 0.4.9)
ggpubr (version 0.4.0)
reshape2 (version 1.4.4)
tibble (version 3.0.4)
gtable (version 0.3.0)
RColorBrewer (version 1.1-2)
plyr (version 1.8.6)
ggrepel (version 0.8.2)
GenomicRanges (version 1.38.0)
rlist (version 0.4.6.2)
tidytext (version 0.2.3)
stringr (version 1.4.0)
data.table (version 1.13.2)
DiagrammR (version 1.0.1)
magrittr (version 2.0.1)
BSgenome.Hsapiens.UCSC.hg19 (version 1.4.0)
BSgenome.Mmusculus.UCSC.mm10 (version 1.4.0)
deconstructSigs (version 1.8.0)
QuPath (0.2.0)
Fiji (.2.0.0)
FloJo (10.8.1)
Prism (9.4.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

PEACE and BDRE Duplex-seq: EGAS00001006951
Duplex-seq data generated from PEACE study samples during this study are not publicly available and restrictions apply to the availability of these data. Such Duplex-seq data are available through the Cancer Research UK & University College London Cancer Trials Centre (ctc.peace@ucl.ac.uk) for academic, non-commercial research purposes upon reasonable request, and subject to review of a project proposal that will be evaluated by a PEACE data access committee, entering into an appropriate data access agreement and subject to any applicable ethical approvals.

Duplex-seq data generated from the BDRE study are available through 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.

The Duplex-seq data for the BDRE and PEACE studies were generated using a larger panel of probes that covered ~50 kb of the genome, spanning hotspots frequently mutated in cancers. This full data set has been provided for the 17 never smokers from the PEACE study. For all other samples, only data for the EGFR and KRAS regions queried are included in this manuscript.

COPA Study RNASeq: EGAS00001006966
De-identified participant data are available upon reasonable request to Dr. Chris Carlsten (christopher.carlsten@ubc.ca) for academic, non-commercial research purposes. Data availability is subject to a data access agreement and applicable ethical approvals.

Mouse WGS: PRJEB58221/ERP143287
Mouse RNA-seq: PRJEB59269/ERP144330

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

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size of 421 patients represents the half-way point of the TRACERx longitudinal study. In total 432 tumours (1553 tumour regions) of the 421 patients were analyses in this study.
Data exclusions	Please see study inclusion/exclusion criteria below. Additionally, samples which fail quality control metrics including low tumor purity (<10%) were also excluded from analysis.
Replication	This is the half-way point of the TRACERx 421 and reflects hypothesis generating analysis.
Randomization	No randomization was conducted
Blinding	Not applicable for this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

MHC-II-FITC Biolegend 107605 M5/114.15.2
CD45-PerCP Biolegend 103130 30-F11
PD-L1-APC Biolegend 124312 10F.9G2
Ly6C-BV421 Biolegend 128032 HK1.4
CD24-BV605 Biolegend 101827 M1/69
CD206-BV711 Biolegend 141727 C068C2
CD86-BV785 Biolegend 105043 GL-1
CD11c-BUV395 BD Biosciences 564080 HL3
Ly6G-BUV661 BD Biosciences 741587 1A8
CD11b-BUV737 BD Biosciences 564443 M1/70
CD64-Pe-Cy7 Biolegend 139314 X54-5/7.1
CD64-APC Biolegend 139306 X54-5/7.1
CD326 (EpCAM)-APC-Fire750 Biolegend 118208 G8.8
Ter119-BV421 Biolegend 116233 TER-119
CD45-BV421 Biolegend 103133 30-F11
CD45-PerCP-Cy5.5 Biolegend 103132 30-F11
CD31-BV421 Biolegend 102423 390
CD49f-PE-Cy7 eBioscience 25-0495-82 GoH3
SiglecF-PE BD Biosciences 552126 E50-2440
Immunohistochemistry/Immunofluorescence
EGFR L858R CST 3197S 43B2
CD68 Abcam ab283654 EPR23917-164
RFP Rockland 600-401-379 Polyclonal
CD11b Abcam ab133357 EPR1344

Validation

The antibodies used have been validated accordingly to manufacturer's instructions. Spleen was used to validate flow antibodies and positive control sections were used to validate all antibodies for IF or IHC staining.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	We used both male and female mice between 6 and 15 weeks of age. All mice were bred in house at the Francis Crick Biological Research Facility, according to UK Home Office Regulations. Light cycles were 7am-7pm, the ambient temperature and humidity was maintained at 21 degrees and 50%, respectively.
Wild animals	Did not involve wild animals
Field-collected samples	Did not involve field collected samples
Ethics oversight	All experiments in this study were approved by the Francis Crick Institute review committees, and conducted according to UK Home Office regulations under project license PPL P8AA77917 & PP6089376

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Please note that the study started recruiting patients in 2016, when TNM version 7 was standard of care. The up-to-date inclusion/exclusion criteria now utilizes TNM version 8. TRACERx inclusion and exclusion criteria Inclusion Criteria: _Written Informed consent _Patients ≥18 years of age, with early stage I-IIIB disease (according to TNM 8th edition) who are eligible for primary surgery. _Histopathologically confirmed NSCLC, or a strong suspicion of cancer on lung imaging necessitating surgery (e.g. diagnosis determined from frozen section in theatre) _Primary surgery in keeping with NICE guidelines planned _Agreement to be followed up at a TRACERx site _Performance status 0 or 1 _Minimum tumor diameter at least 15mm to allow for sampling of at least two tumour regions (if 15mm, a high likelihood of nodal involvement on pre-operative imaging required to meet eligibility according to stage, i.e. T1N1-3) Exclusion Criteria: _Any other* malignancy diagnosed or relapsed at any time, which is currently being treated (including by hormonal therapy). _Any other* current malignancy or malignancy diagnosed or relapsed within the past 3 years**. *Exceptions are: non-melanomatous skin cancer, stage 0 melanoma in situ, and in situ cervical cancer **An exception will be made for malignancies diagnosed or relapsed more than 2, but less than 3, years ago only if a pre-operative biopsy of the lung lesion has confirmed a diagnosis of NSCLC. _Psychological condition that would preclude informed consent _Treatment with neo-adjuvant therapy for current lung malignancy deemed necessary _Post-surgery stage IV _Known Human Immunodeficiency Virus (HIV), Hepatitis B Virus (HBV), Hepatitis C Virus (HCV) or syphilis infection. _Sufficient tissue, i.e. a minimum of two tumor regions, is unlikely to be obtained for the study based on pre-operative imaging Patient ineligibility following registration _There is insufficient tissue _The patient is unable to comply with protocol requirements _There is a change in histology from NSCLC following surgery, or NSCLC is not confirmed during or after surgery. _Change in staging to IIIC or IV following surgery _The operative criteria are not met (e.g. incomplete resection with macroscopic residual tumors (R2)). Patients with microscopic residual tumors (R1) are eligible and should remain in the study _Adjuvant therapy other than platinum-based chemotherapy and/or radiotherapy is administered. PEACE Patient characteristics: Inclusion criteria · Age 18 years or older · Confirmed diagnosis of any form of solid malignancy with metastatic disease (where the site of origin is known or unknown), with the exception of primary brain tumour in which there may not be evidence of metastatic disease · Oral and written informed consent from patient to enter the study and to undergo tissue harvesting after death or informed consent from a nominated representative or a person in a qualifying relationship after the patient has died. Exclusion Criteria · Medical or psychiatric condition that would preclude informed consent · History of intravenous drug abuse within the last 5 years · Confirmed diagnosis of known high-risk infections (e.g. HIV/AIDS-positive, hepatitis B/C, tuberculosis and Creutzfeldt-Jacob disease) unless patient case is of a particular scientific interest and agreed in advance with local mortuary staff and pathologist.

Recruitment	TRACERx: Patients seen with a new diagnosis of lung cancer in lung cancer units across the United Kingdom, according to the eligibility criteria above, were recruited. No selection bias has been identified to date. PEACE Recruitment: patients seen with metastatic disease of solid malignancy across England and Scotland, according to the eligibility criteria above, were recruited. BDRE: For recruitment, patients undergoing bronchoscopy for an indeterminate pulmonary nodule were recruited from the Chest Clinics at the Rocky Mountain Regional VA Medical Center.
Ethics oversight	The TRACERx study was approved by the NRES Committee London with the following details: Study title: TRAcking non small cell lung Cancer Evolution through therapy (Rx) REC reference: 13/LO/1546 Protocol number: UCL/12/0279 IRAS project ID: 138871 The PEACE study was approved by the NRES Committee London with the following details: Study title: The PEACE (Posthumous Evaluation of Advanced Cancer Environment) Study REC reference: 13/LO/0972 IRAS project ID: 125424 COMIRB approved protocol (00-1108) "Biomarkers and Dysplastic Respiratory Epithelium"

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	TRACERx Lung https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent Research Ethics Committee, 13/LO/1546 PEACE https://clinicaltrials.gov/ct2/show/NCT03004755, approved by independent Research Ethics Committee,13/LO/0972 Biomarkers and Dysplastic Respiratory Epithelium https://clinicaltrials.gov/ct2/show/NCT00900419
Study protocol	https://clinicaltrials.gov/ct2/show/NCT01888601 https://clinicaltrials.gov/ct2/show/NCT03004755 https://clinicaltrials.gov/ct2/show/NCT00900419
Data collection	TRACERx & PEACE: Clinical and pathological data is collected from patients during study follow up - this period is a minimum of five years. Data collection is overseen by the sponsor of the study (Cancer Research UK & UCL Cancer Trials Centre) and takes place in hospitals across the United Kingdom. A centralised database called MACRO is used for this purpose.
Outcomes	TRACERx and PEACE: Disease-free outcome data were gathered for all patients. Last updated in 15 June 2021

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For flow cytometry analysis of immune cells, mouse lungs were minced into small pieces, incubated with collagenase (1 mg/ml; ThermoFisher) and DNase I (50 U/ml; Life Technologies) for 45 min at 37°C and filtered through 100 µm strainers (Falcon). Red blood cells were lysed for 5 min using ACK buffer (Life Technologies). Cells were stained with fixable viability dye eFluor780 (BD Horizon) for 30 min and blocked with CD16/32 antibody (Biolegend) for 10 min. Cells were then stained with antibodies for 30 min (see Supplementary Table S10). Intracellular staining was performed using the Fixation/Permeabilization kit (eBioscience) according to the manufacturer's instructions. Samples were resuspended in FACS buffer (2% FCS in PBS). For flow cytometry sorting of AT261, epithelial and immune cells, minced lung tissue was digested with Liberase TM and TH (Roche Diagnostics) and DNase I (Merck Sigma-Aldrich) in HBSS for 30min at 37°C in a shaker at 180rpm. Samples were passed through a 100µm filter, centrifuged (300 x g, 5 min, 4 degrees) and red blood cells lysed as above. Extracellular

	antibody staining was then performed followed by incubation in DAPI (Sigma Aldrich) to label dead cells. Gating strategies for sorting and analysis are outlined in Extended Data Figure 6.
Instrument	Flow cytometry analyses were carried out on a BD Fortessa Symphony A5. All cell-sorting experiments were carried out on a BD Influx cell sorter, Aria Fusion or Aria III.
Software	FlowJo10.4.2 (FlowJO,LCC 2006-2018, USA) was used for analysis
Cell population abundance	Purity check was routinely performed after each sorting. Cells were used when purity was above 85%. flow cytometry data lineages are described as % of CD45 and phenotypes are described as % of parent. In all cases, populations are indicated on figures and in figure legends
Gating strategy	All samples were first gated to exclude cellular debris using SSC-A/FSC-A and a diagonal gate using FSC-H/FSC-A was then used to exclude doublets. Live cell discrimination was next determined using viability dyes. All gating strategies are defined in Extended Data Figure 6A-C

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.