nature portfolio

Corresponding author(s): Charles Swanton
Julian Downward
George Kassiotis

Last updated by author(s): Dec 18, 2022

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	FACSDiva v8.0 (BD) Everest v2.4 (BioRad)
Data analysis	FlowJo v10 (Tree Star Inc.) Excel 2016 (Microsoft) SigmaPlot 14.0 (Systat Software Inc.) GraphPad Prism 7 (GraphPad Software) Qlucore Omics Explorer v3.3 (Qlucore) Scaffold software (Proteome Software Inc.) Salmon v0.12.0 GNU parallel R (versions 3.6.1-4.0.0) TRUST4 v1.0.8 CellRanger V(D) (10x Genomics) Imaris software 9.8 (Bitplane) The packages dplyr (v1.0.7), data.table (v1.14.2), tidyverse (v 1.3.1) and rjson (v0.2.20) were used for data handling in R. Statistical analysis in R: The package Hmisc (v 4.6.0) was used for Spearman's correlation analysis. The package lme4 (v1.1.27.1) was used for linear mixed effects models. The package survival (v3.2.13) was used for statistical associations with patient outcome metrics. UCSC Xena browser (https://xena.ucsc.edu)

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

The RNA sequencing (RNA-seq) and whole exome sequencing (WES) data (in each case from the TRACERx study) used during this study have been deposited at the European Genome–phenome Archive (EGA), which is hosted by The European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG) under the accession codes EGAS00001006517 (RNAseq) and EGAS00001006494 (WES); access is controlled by the TRACERx data access committee. Details on how to apply for access are available at the linked page. Other data supporting the findings of this study are available within the paper and its supporting information files, with raw data openly available from the Francis Crick Institute in a Figshare repository (https://crick.figshare.com). TCGA and GTEx data used for the analyses described in this manuscript were obtained from dbGaP (https://dbgap.ncbi.nlm.nih.gov) accession numbers phs000178.v10.p8.c1 and phs000424.v7.p2.c1 in 2017. Additional TCGA LUAD expression data and average copy number of the ERVK-7 genomic location data were downloaded from the UCSC Xena browser (https://xena.ucsc.edu). Nucleotide sequences were downloaded from NCBI nucleotide resources (https://www.ncbi.nlm.nih.gov/nuccore). Source data are provided with this paper.

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	No statistical methods were used to predetermine sample size. The final target of the TRACERx study is to recruit a cohort of 842 patients, required to detect at least a 23% relative risk reduction and a 10% improvement in 5 -year overall survival. The data used here represents the half-point of this study. For in vivo mouse experiments, group sizes were determined based on prior experience with the respective models and the results of our preliminary experiments. The number of repeats were determined by the balance between statistical significance and reduction in animal use.
Data exclusions	No data were excluded.
Replication	TRACERx is a prospective longitudinal study. As such, the results shown here are not the result of an experimental setup. This is the half-way point of the TRACERx 421 and reflects hypothesis generating analysis. Findings from TRACERx were validated using independent cohorts. Experiments involving animals were repeated multiple times, as indicated in the figure legends, and all attempts at replication were successful.
Randomization	Given the descriptive nature of the TRACERx longitudinal study, no experimental groups were allocated beforehand. For in vivo mouse studies, mice were randomly allocated to the different treatment groups, with the exception of specific genotypes that were allocated to different groups according to genotype.
Blinding	For in vivo mouse studies, investigators were not blinded to group allocation in experimental setup, data collection or analysis. Blinding was not required as data were based on quantitative analysis of phenotypes.

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

In vivo
PDL1 (200ug 10F.9G2; BioXCell BE0101)
CTLA4 (200ug 9H10; BioXCell BE0131)
CXCL13 (200ug 143614; R&D Systems MAB470)
NK1.1 (200ug PK136; BioXCell BE0036)
CD8 (200ug 53-6.7; BioXCell BE0004-1)
eMLV env (200ug 83A25; in-house)
KARV env (200ug J1KK; in-house)

Flow cytometry
CD45 (1:200 30-F11; Biolegend 103111)
B220 (1:200 RA3-6B2; Biolegend 103207)
GL7 (1:200 GL7; Biolegend 144603)
CD95 (1:200 SA362F7; Biolegend 152617)
CXCR4 (1:200 L276F12; Biolegend 146511)
CD86 (1:200 GL-1; Biolegend 105043)
TCRB (1:200 H57-597; Biolegend 109225)
CD4 (1:200 GK1.5; Biolegend 100431)
PD1 (1:200 29F.1A12; Biolegend 135209)
CXCR5 (1:200 L138D7; Biolegend 145503)
anti-mouse IgG (1:200 Poly4060; Biolegend 406001)
anti-mouse IgA (1:200 11-44-2; Southern Biotech 1165-02)
anti-mouse IgM (1:200 RMM-1; Biolegend 406517)
anti-human IgG (1:200 M1310G05; Biolegend 410703)
anti-human IgA (1:200 130-114-002; Miltenyi Biotech 130-113-476)
anti-human IgM (1:200 MHM-88; Biolegend 314510)

Immunohistochemistry
B220 (1:250 RA3-6B2; BD Biosciences 553086)
CD8 (1:250 4SM15; Thermo Fisher 14-0808-82)
Ki67 (1:250 MIB-1; Agilent M7240)
NCR1 (1:250 EPR23097-35; Abcam ab233558)
PNA (1:250; Vector Biolaboratories B-1075)
ERVK-7 (1:250 polyclonal; Thermo Fisher PA5-49515)
HRP anti-rat IgG (1:1000 polyclonal; Thermo Fisher 31470)
HRP anti-mouse IgG (1:1000 polyclonal; Thermo Fisher 31430)
HRP anti-rabbit IgG (1:1000 polyclonal; Thermo Fisher A16116)

2D and 3D immunofluorescence
CD3 (1:100 polyclonal; Abcam ab5690)
B220 (1:100 RA3-6B2; Biolegend 14-0452-82)
TTF (1:100 8G7G3/1; Abcam ab72876)
Alexa Fluor 546 anti-rabbit IgG (1:100 polyclonal; Thermo Fisher A-10040)
Alexa Fluor 546 anti-rabbit IgG (1:200 polyclonal; Thermo Fisher A-11035)
Alexa Fluor 594 anti-rabbit IgG (1:100 polyclonal; Thermo Fisher R37119)
Alexa Fluor 488 anti-rat IgG (1:100 polyclonal; Thermo Fisher A-21208)
Alexa Fluor 488 anti-rat IgG (1:200 polyclonal; Thermo Fisher A-11006)
Alexa Fluor 647 anti-rat IgG (1:100 polyclonal; Thermo Fisher A-48272)
Alexa Fluor 488 anti-mouse IgG (1:100 polyclonal; Thermo.Fisher A-21202)
Alexa Fluor 647 anti-goat IgG (1:100 polyclonal; Thermo Fisher A-21447)

Validation

Validation data of all commercial antibodies are available on vendor websites and antibody datasheets. Specificity has been validated by staining for the immunogen (flow cytometry, immunofluorescence or Western blotting) and have been used extensively in numerous other studies.

For the ERVK-7 antibody in particular (Thermo Fisher PA5-49515), cross-reactivity against other members of the HERV-K(HML-2) family has not been examined by the vendors. Based on sequence conservation among HERV-K(HML-2) members of the part of the

envelope glycoprotein that was uses as the immunogen, it is highly likely that this polyclonal antibody reacts with several members. We therefore refer to it in the manuscript as HERV-K(HML-2)-reactive.

The specificity of the newly generated J1KK antibody was established by staining cells lines expressing or not expressing the target antigen. These results are shown in Fig. 2i.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	KPB6 (Francis Crick Institute Cell Services) KPAR1.3 (in-house) KPAR1.3<G12C> (in-house) HEK293T cells (Francis Crick Institute Cell Services) EL4 (Francis Crick Institute Cell Services) CTLL2 (Francis Crick Institute Cell Services) B16 (Francis Crick Institute Cell Services) 4T1 (Francis Crick Institute Cell Services) 3LL (Francis Crick Institute Cell Services) MC38 (Francis Crick Institute Cell Services) A549 (Francis Crick Institute Cell Services) HBEC (Francis Crick Institute Cell Services) NK92 (Francis Crick Institute Cell Services) Mus dunni (Francis Crick Institute Cell Services) HEK293T.ERV3-1env (in-house) HEK293T.HERV-K(HML-2)env (in-house)
Authentication	DNA fingerprinting for human cells lines
Mycoplasma contamination	Verified as mycoplasma-free
Commonly misidentified lines (See ICLAC register)	Although not commonly misidentified, there is some ambiguity as to the origin of EL4 cells. In contrast to human cells, murine cell line authentication by DNA fingerprinting is not yet established and it is therefore difficult to know which EL4 subline might be closer to the original. We have chosen to use the EL4 cells at the Francis Crick Institute as they are the only variant that we find to be negative for infectious MLVs.

Animals and other organisms

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

Laboratory animals	C57BL/6J wild-type Aicdatm1.1(cre/ERT2)Crey (AicdaCreERT2) Ighg1tm1(cre)Cgn (Ighg1Cre) Gt(ROSA)26Sortm1(EYFP)Cos (Rosa26LSL-EYFP) Gt(ROSA)26Sortm1(CAG-Brainbow2.1)Cle (Rosa26LSL-Confetti) Emv2-deficient mice Mice were housed in ventilated cages kept in constant temperature (21-25°C) and humidity (50-60%), with standard 12-hour light/dark cycles, and under specific pathogen-free conditions.. 8 to 12-week-old male or female mice were used for all experiments.
Wild animals	No wild animals were used in the study.
Field-collected samples	No field collected samples were used in the study
Ethics oversight	All experiments were approved by the ethics committee of the Francis Crick Institute and conducted according to local guidelines and UK Home Office regulations under the Animals Scientific Procedures Act 1986 (ASPA).

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	421 patients are included in this TRACERx cohort. 44.6% are females , 55.4% males; 93% are smokers of have a smoking history, 7% are never smokers; 25% of patients were diagnosed at stage IA, 25% at IB, 17.8% at IIA, 13.5% at IIB, 18.5% at IIIA and 0.2% at IIIB; 52% of diagnosed tumours were adenocarcinomas, 28.8% were squamous cell carcinomas and 19.2% were of other histological subtypes; 93% of the cohort is from a white ethnic background and the mean age of the patients is 69, ranging between 34 and 92. 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.

Recruitment

When patients are initially diagnosed with stage I-III lung cancer and then referred for surgical resection, a research nurse identifies them on a clinic/operating list. The patient has an initial eligibility assessment and then provided with written information about the TRACERx study and he/she can ask the research nurse any questions.

Patients have to agree to provide serial blood samples whenever they attend clinic for routine blood sampling, so this represents the only main potential self-selecting bias (i.e. only patients willing to do this would participate). However, it is unclear how this would affect the biomarker analyses. Also, the gender and ethnicity characteristics are in line with patients seen in routine practice.

Inclusion and exclusion criteria are summarised above.

All patients were assigned a study ID that was known to the patient. These were subsequently converted to linked study IDs such that the patients could not identify themselves in study publications. All human samples, tissue and blood, were linked to the study ID and barcoded such that they were anonymised and tracked on a centralised database overseen by the study sponsor only. Written informed consent was obtained from all patients.

Ethics oversight

The study was approved by the National Research Ethics Service (NRES) Committee London - Camden and Islington, with sponsor's approval of the study by University College London (UCL) 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

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: https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent Research Ethics Committee, 13/LO/1546

Study protocol

TRACERx: https://clinicaltrials.gov/ct2/show/NCT01888601

Data collection

Recruitment commenced April 2014. Clinical and pathological data are collected from patients for a minimum of five years. Study co-ordination and data collection are overseen by the study sponsor (Cancer Research UK & UCL Cancer Trials Centre). A centralised database with remote data entry (MACRO) was used. Patients were recruited from London, Leicester, Manchester, Aberdeen, Birmingham, and Cardiff. Recruitment was completed at all sites on December 16, 2021 except at London and Manchester hospital sites where recruitment is due to complete March 31, 2022.

Outcomes	TRACERx: Disease-free survival (DFS) is measured from the time of study registration to date of first lung recurrence or death from any cause. Patients who do not have these events are censored at the date last known to be alive (including patients who developed a new primary tumour that has been shown biologically to not be linked to the initial primary lung tumour). TCGA: Overall survival (OS) is the time from study registration until death from any cause. For both DFS and OS, patients without an event are censored at the date they were last known to be alive (and also recurrence-free for DFS). TRACERx primary outcome: determine the clinical impact of intratumour heterogeneity on the clinical course of disease and the impact of adjuvant platinum-based chemo on intratumour heterogeneity in relapsed disease. TRACERx secondary outcome: No secondary outcome was pre-defined

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	Lungs were perfused with 20 mL cold PBS, cut into small pieces, and incubated with 1 mg/mL collagenase (Thermo Fisher) and 50 U/mL DNase I (Life Technologies) in PBS for 30 mins at 37oC. Single-cell suspensions were prepared from the spleens or lymph nodes by mechanical disruption. Samples were filtered through 70 µm nylon strainers and red blood cells were lysed using 0.83% ammonium chloride before resuspension in FACS buffer (PBS, 2% FCS, 0.05% sodium azide). Cell lines were grown under standard conditions.
Instrument	Samples were run on a LSR Fortessa or a Ze5 analyser
Software	Samples were run on a LSR Fortessa running BD FACSDiva v8.0 or a Ze5 analyser running BioRad Everest v2.4 and analysed with FlowJo v10.
Cell population abundance	Sorted B cells were >95% pure and purity was confirmed additionally by subsequent scRNA-seq.
Gating strategy	For the identification of GC B cells and Tfh cells, cell suspensions were first gated on FSC-A and SSC-A, following by FSC-A and FSC-H to discriminate single cells from doublets. Live cells were identified by gating on NIR Live/Dead staining, and immune cells within live cells by gating on CD45+ cells. GC B cells were gated as B220+ first, following by CD95+ GL7+ double-positive gating. Tfh cells were gated as CD4+ TCRb+ double-positive first, following by CXCR5+ PD-1+ double-positive gating. Examples of these gating strategies is shown in Extended Data Fig. 12a. For antibody assays, HEK293T sublines were first gated on FSC-A and SSC-A, following by FSC-A and FSC-H to discriminate single cells from doublets. HEK293T sublines were then discriminated based on the intensity of GFP expression. Serum antibody binding was assessed by the increase in the intensity of staining with the respective secondary antibody. Examples of these gating strategies is shown in Extended Data Fig. 12b.

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