nature research

Corresponding author(s): Peter Campbell

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Image processing from sequencing data using standard Illumina X10 pipeline
Data analysis	Alignment and variant calling performed using Sanger Institute's custom pipeline. Single-nucleotide substitutions were called using the CaVEMan (cancer variants through expectation maximization) algorithm (https://github.com/cancerit/CaVEMan). Small insertions and deletions were called using the Pindel algorithm (https://github.com/genome/pindel). Rearrangements were called using the BRASS (breakpoint via assembly) algorithm (https://github.com/cancerit/BRASS). List of programs and softwares: R – version 3.5.1 BWA-MEM - version 0.7.17-r1188 (https://sourceforge.net/projects/bio-bwa/) CaVEMan - version 1.11.2 Pindel - version 2.2.5 Brass - version 6.1.2 ASCAT NGS - version 4.1.2 Xenome (https://github.com/data61/gossamer/blob/master/docs/xenome.md) deepSNV - version 1.28.0 (https://bioconductor.org/packages/release/bioc/html/deepSNV.html) ANNOVAR (http://wannovar.wglab.org/) IGV (http://software.broadinstitute.org/software/igv/) JBrowse (https://jbrowse.org/)

cgpVAF (https://github.com/cancerit/vafCorrect)
RPhylip - version 0.1.23 (http://www.phytools.org/Rphylip/)
hdp - version 0.1.5 (https://github.com/nicolaroberts/hdp)
MutationalPatterns - version 1.8.0 (https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html)
dNdScv - version 0.0.1 (https://github.com/im3sanger/dndscv)
Telomerecat - version 3.1.2 (https://github.com/jhrf/telomerecat)

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data that support the findings of this study have been deposited in the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/home) under accession number EGAD00001005193. Somatic mutation calls, including single base substitutions, indels and structural variants, from all 632 samples have been deposited on Mendeley Data with the identifier: http://dx.doi.org/10.17632/b53h2kwpyy.2.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	Sample size was chosen to give good representation of inter-patient and intra-patient variability in mutation burden.
Data exclusions	Oligoclonal colonies or with low mean coverage (<8x) (Extended Figure 2e) were excluded due to the inaccuracy of mutation catalogues. One outlying cell in an ex-smoker with >10,000 mutations is excluded from the plot of Figure 4 to improve visualisation.
Replication	No experimental replication has yet been attempted.
Randomization	Not applicable - this is a descriptive study, not an intervention study.
Blinding	Not applicable - all dependent variables were computationally generated (mutation counts, signatures etc) and statistical analyses were prespecified.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

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

Antibodies

Antibodies used	Antibodies used in this study were described in OnlineMethods.

Validation	Antibodies were validated by the manufacturer.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	3T3-J2 feeder cells were kindly provided by Prof. Fiona Watt (King's College London)
Authentication	The feeder cells had whole genome sequencing performed, confirming their murine origin and clonal derivation.
Mycoplasma contamination	They were tested negative for Mycoplasma by PCR test (PMCID: PMC202165)
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics	We analysed single cell-derived colonies from bronchial epithelium of 16 subjects, including 3 children, 4 never-smokers, 6 ex-smokers and 3 current smokers. Clinical characteristics of the cohort are described in Supplementary Table 1. Of the ex-smokers, 2 had had a previous cancer treated with curative intent, and 5 had a carcinoma in situ or invasive squamous cell carcinoma.The children in the cohort had bronchoscopy for investigation or follow-up of congenital anomalies.
Recruitment	Recruited through University College Hospitals, London, UK. Our cohort does potentially suffer from recruitment bias, since samples could only ethically be obtained from individuals undergoing a clinically indicated bronchoscopy.

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	The epithelium was dissected away from the underlying stroma and fetal bovine serum (FBS) was added to a final concentration of 10%. Both the epithelium and stroma were combined and digested in 0.1% trypsin/EDTA at 37□C for 30 minutes. The solution was neutralised with FBS to a final concentration of 10% and added to the neutralised dispase solution1. Cells were passed through a 100 □m cell strainer and stained in sorting buffer (1x PBS, 1% FBS, 25 mM HEPES and 1 mM EDTA) with anti-CD45-PE (BD Pharminogen 555483, 1:200), anti-CD31-PE (BD Pharminogen 555446, 1:200), anti-EPCAM-APC (Biolegend 324208, 1:50) antibodies and DAPI (1 ug/ml).
Instrument	BD FACSAria Fusion
Software	Supplementary Methods
Cell population abundance	Supplementary Methods
Gating strategy	Supplementary Methods

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