nature research

Corresponding author(s): Prof. Sam Janes

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	We used the GenomeStudio Gene Expression Module v1.0 software (Illumina) to normalize the DASL gene expression raw data. Primer Express® Software v2.0 (Thermo Fisher Scientific) was used to design gene-specific primers used for qPCR validation.
Data analysis	Data analysis was performed using R version 3.5.0 and Bioconductor version 3.7. All analysis code is stored in a Github repository at github.com/ucl-respiratory/preinvasive (currently private, to be made public on publication). The version associated with this submission is tagged as 2.1, a copy of which has been made available to reviewers. All software package versions for Bioconductor dependencies are included as a file in this Github repository.

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

Whole-genome sequencing data have been deposited at the European Genome Phenome Archive (https://www.ebi.ac.uk/ega/ at the EBI) with accession number EGAD00001003883. Due to the identifiable nature of these data, researchers must submit a data access request through the EGA to obtain access. All gene expression and methylation microarray data reported in this study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) public repository, and they are openly accessible through GEO accession number GSE108124.

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	Our study include a unique collection of high-grade pre-invasive lung lesions therefore we used all human pre-invasive samples that were available to determine the genomic, epigenomic and transcriptomic profiles of these lesions.
Data exclusions	CIS samples with less than 200ng of RNA and 1ug of DNA were excluded from the study as they were not suitable for DASL gene expression and methylation microarrays, respectively. For DNA, only samples with an A260/280 ratio of 1.7-1.9 were included in the study. Otherwise there were no systematic data exclusions.
Replication	We used qPCR to validate the gene expression of eight genes. We observed a high degree of correlation (r=0.982) between qPCR and microarray data. The lung pre-invasive biopsies are extremely small so it was not possible to perform technical replicates.
Randomization	Methylation and gene expression samples were divided into training/validation sets randomly, before profiling was done. The investigators balanced clinical outcomes between the groups to allow similar numbers of progressive/regressive samples, but had no prior knowledge of their molecular profiles.
Blinding	This study is observational, therefore blinding is not relevant as no intervention was performed.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patient characteristics are summarised in Table 1. Raw data is available from the NCBI GEO database, as described in the Methods section. In comparing progressive and regressive samples, we found that progressive samples were associated with a higher pack-year smoking history in the methylation discovery group only ($p < 0.01$) and with increased age in the WGS group (p

= 0.01). No clinical differences were consistently observed across the different analysis groups.

Recruitment

All patients with pre-invasive lung cancer lesions were recruited through University College London Hospitals (UCLH) Early Lung Cancer Surveillance Programme (ELCSP). Full details of the surveillance protocol including eligibility criteria for patient inclusion have been previously described (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2111271/). Due to a higher chancer of detecting preinvasive lesions, patients with a history of lung cancer are overrepresented in this cohort compared to the general population; however we have controlled for this factor and do not believe it significantly affects our conclusions.