nature portfolio

Corresponding author(s): Janne Lehtiö

Last updated by author(s): Jun 24, 2021

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

-Proteomics Data

Data-Dependent-Acquisition (DDA):
proteowizard from msConvert (v.3.0.19127), MSGF+ (v2017.07.21), Percolator (v3.1), Nextflow pipeline (https://github.com/lehtiolab/nf-workflows, commit: 898bb20)
OpenMS project's IsobaricAnalyzer (v2.0)

Data-Independent-Acquisition (DIA):
Spectronaut (version 13.10) from Biognosys, Pulsar

For further details, please see the methods section.

Data analysis

- R version 3.6.2 or higher with packages:
survminer (v0.4.8),
survival (v3.2-7),
DEqMS(v1.6.0),
IlluminaHumanMethylation450kprobe (v2.0.6)
TxDb.Hsapiens.UCSC.hg19.knownGene (v3.2.2)
GenomicRanges (v1.34.0),
ConsensusClusterPlus(v1.50.0)
NMF(v.0.23.0)
ComplexHeatmap (v2.2.0)
ggplot2 (v3.3.3),
Seurat(v4.0.0),

PCAtools(v1.2.0),
clusterProfiler(v3.14.3),
ESTIMATE (v1.0.11),
GSVA (v1.34.0),
biomaRt (v2.44.1)
switchbox(v1.24.0)
dunn.test(v1.3.5)

- ANNOVAR (2020Apr01)

- Python3
scikit-learn (v0.21.2)
numpy (v1.17.4)

- Illumina bcl2fastq2 Conversion Software v2.20

- BALSAMIC workflow v4.0.067.
- FastQC v0.11.5
- fastp v0.20.0
- BWA MEM v0.7.15
-samtools v1.6
-MarkDuplicate v2.17.0
-MultiQC v1.7
-VarDict v2019.06.04
-Ensembl VEP v94.5

- Molecular Tumour Board Portal (accessed 2/2020, Tamborero D., Nature Medicine. July 2020))

- IPAW workflow run in Python 3, powered by Nextflow and containerized in Docker (v0.2)

- ImageJ (v. 1.53f)

Details and references can be found within text in the relevant Methods sections.

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 mass spectrometry proteomics data for DDA and DIA analyses have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD020191 (DDA discovery cohort), PXD020548 (DIA discovery and late-stage cohorts), and PXD025560 (DIA validation cohort).
For panel sequencing, sequence data has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001005482.
Previously published proteomics data that was re-analysed in this study are available in PRIDE with the identifier PXD010429, in iProx Consortium with the subproject ID IPX0001804000 and CPTAC Data Portal (https://cptac-data-portal.georgetown.edu/study-summary/S056).
Previously published gene expression data that were re-analysed here are available under accession codes GSE60645 and GSE149521, and in ArrayExpress with the identifier E-MTAB-6043.
The human [Pan-Cancer Atlas and lung adenocarcinoma (LUAD) gene expression data] data were derived from the TCGA Research Network: http://cancergenome.nih.gov/. The data-set derived from this resource that supports the findings of this study is available at https://gdc.cancer.gov/access-data.
Previously published resource of drug sensitivity in cancer cell lines data are available at https://www.cancerrxgene.org/ .
Source data for all figures and Extended Data figures have been provided as Source Data files. All other data supporting the findings of this study are available from the corresponding author upon reasonable request.

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	Tumor analysis: This was an observational study, thus no statistical methods used to predetermine to sample size. Statistical significance was evaluated based on the observed group sizes and corrected for multiple testing. Resected lung cancer tumour samples from a total of 192 patients with early-stage lung cancer surgically treated at the Skåne University Hospital in Lund, Sweden, were provided and collected. Mechanistic cell-based study: 3 cell culture replicates were performed for each of the experiments and all three were presented in the paper.
Data exclusions	For the current proteomics analysis, 51 samples out of 192 were excluded based on pre-established criteria listed below. 35 samples were excluded due to insufficient protein amount or deviating Protein-RNA or Protein-DNA concentration correlation resulting in 157 samples remaining for protein digestion and further MS analysis. Four out of 157 samples had insufficient peptide amount (< 100 µg) for TMT labeling and were excluded. All remaining 153 samples were pre-screened by LC-MS/MS on a QExactive HF using short gradient (60 min) DDA runs to identify outlier samples. Based on analysis of the short gradient data, 10 samples with extensive blood contamination were excluded, resulting in 143 samples remaining for tandem mass tag (TMT) labeling. Subsequent re-analysis of clinical data resulted in the exclusion of two additional samples after MS data generation due to uncertain primary tumour origin. This resulted in a final cohort size of 141 lung cancer samples for subsequent analysis.
Replication	Tumor analysis: Two technical replicates were performed for 6 tumor samples, each. This data was included in the analysis and verified robustness of the measurements. This study also included two additional cohorts on which the SVM and k-TSP classifiers were tested and validated. These comprised 84 and 209 samples, 1 sample was excluded from the latter cohort due to insufficient material for analysis. The validation criteria included histology, protein marker, mutation signature, and other phenotypic characteristics associated with the identified subtypes. Mechanistic cell-based study: experiments were performed on three different cell lines (2 lung cancer and 1 liver cancer, all results are presented in the paper), none of the results contradicted each other.
Randomization	Not relevant to our study because this was an observational study.
Blinding	Immunohistochemistry (IHC) scoring and proteomics sample preparation and analysis was performed in a blinded manner to avoid observation bias. Meta data for the cohorts was consulted only after the downstream analysis.

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	PDL1 (CD274): anti-PD-L1 antibody (rabbit monoclonal antibody clone 28-8, dilution 1:100, ab205921, Abcam, UK) and employing an OptiView DAB IHC Detection kit (cat. No. 760-700, Roche Diagnostics, Switzerland). CD3: anti-CD3 (polyclonal rabbit antibody, cat. No. A0452, DAKO, Denmark). CD8: anti-CD8 (mouse monoclonal antibody clone C8/144B, cat. No. M7103, DAKO, Denmark). Anti-FGL1 (mouse monoclonal antibody A-8, cat. No. sc-514057, Lot J0520, Santa Cruz Biotechnology). Anti-LKB1 (STK11) (mouse monoclonal antibody Ley 37D/G6, cat. No. ab15095, Lot GR260328-12, Abcam). Anti-HNF1A (rabbit monoclonal antibody D7Z2Q, cat. No. 89670S, Lot 1, Cell Signaling Technology). Anti-beta-actin (mouse monoclonal antibody C4, cat. No. sc-47778, Lot K1617 Santa Cruz Biotechnology). Anti-alpha-tubulin (mouse monoclonal antibody B7, cat. No. sc-5286, Lot E2010, Santa Cruz Biotechnology). Anti-GAPDH (rabbit polyclonal antibody FL-335, cat. No. sc-25778, Santa Cruz Biotechnology) Anti-HSP90 (monoclonal purified mouse antibody clone 68/Hsp90, cat. No. 610418, BD Biosciences) Mouse anti-rabbit IgG-HRP (cat. No. sc-2357, Lot E0819, Santa Cruz Biotechnology) Sheep anti-mouse IgG-HRP (cat. No. NA931V, Lot 9471754, GE Healthcare UK)
Validation	Antibodies were validated according to manufacturer's information.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	NCI-H1944 (CRL-5907, ATCC) NCI-H1395 (CRL-5868, ATCC) HepG2 (HB-8065, ATCC)
Authentication	HepG2 was authenticated. The NCI-H1944 and NCI-H1395 cell lines were not authenticated due to low passage number after purchasing from ATCC.
Mycoplasma contamination	The cells were tested for Mycoplasma contamination using MycoAlert detection kit purchased from BioNordika (Lonza, cat. No. LT07-118).
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	Patient characteristics and recruitment details for the early-stage cohort have been previously described: Karlsson, A., Brunnström, H., Micke, P., Veerla, S., Mattsson, J., La Fleur, L., Botling, J., Jönsson, M., Reuterswärd, C., Planck, M., Staaf, J., 2017. Gene Expression Profiling of Large Cell Lung Cancer Links Transcriptional Phenotypes to the New Histological WHO 2015 Classification. J Thorac Oncol 12, 1257–1267. The age of patients ranged from 36 to 84 (median 69) years. 53% were female and 47% were male. EGFR, TP53, and KRAS mutations were found in 9%, 68%, and 21% of the patients, respectively. The patients were diagnosed with stage I, II, III, and IV in 66%, 20%, 10% and 4% of cases, respectively. The late-stage cohort comprised biopsy material from inoperable lung cancer patients. The diagnosed stages were I, II, III, and IV in 22%, 18%, 13%, and 39%, respectively, with no staging information available in 8% of the cases. The validation cohort comprised resected lung cancer tumor samples from patients that underwent surgery for lung cancer. 46% of the patients were female and 54% were male. EGFR mutation information was available in 93% of the cases, 12% of which were found positive. TP53 mutation information was available in 82% of the cases, 52% of which were found positive. The patients were diagnosed with stage I, II, III, and IV in 55%, 27%, 15% and 3% of cases, respectively.
Recruitment	All experiments were conducted in agreement with patient consent and ethical review board regulations and decisions. Since this was an observational study, recruitment was driven by availability of the material and was not expected to be affected by biases beyond the staging of the patients (early stage vs late stage), which was considered and described in the study.
Ethics oversight	The study was approved by the Regional Ethical Review Board in Lund, Sweden (Registration no. 2004/762 and 2014/32) and by Regional Ethical Committee for Medical and Health Research Ethics, REK South-East in Oslo, Norway (Ref: S-06402b). All experiments were conducted in agreement with patient consent and ethical review board regulations and decisions.

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