nature research

Corresponding author(s): Xianghong Jasmine Zhou, Edward B. Garon, Wenyuan Li

Last updated by author(s): Dec 16, 2020

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 .

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	sratoolkit 2.9.0 was used to download data under phs001417.v1.p1. No other tools were used to collect data in this study.
Data analysis	All softwares and packages used in our analyses are open-sourced. 1) java 1.8.0_111 2) GATK 3.8.0 3) FLASh2 4) bwa 0.7.17 5) samtools 1.9 6) picard 2.18.4 7) bedtools 2.26.0 8) python 2.7.14 9) numpy 1.13.3 10) pandas 0.20.3 11) scipy 1.1.0 12) sklearn 0.19.1 13) decimal 1.70 14)BAMSurgeon 1.0.0 15)divide-and-conquer algorithm The codes for our algorithm can be found in the code package https://zhoulab.dgsom.ucla.edu/pages/cfSNV, along with a demo and a README description.

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. 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

Data used in this project: dbGaP accession number phs001417.v1.p1, European Nucleotide Archive (ENA) accession number ERS700858, ERS700859, ERS700860, ERS700861, and European Genome-Phenome Archive (EGA) accession number EGAS00001004373. The data at EGA (EGAS00001004373) are available under restricted access, which can be obtained by contacting the corresponding author.
Data under phs001417.v1.p1 are previous published data under controlled access from other groups. These data are used for validation of our method. These data are associated with Fig 2, Fig 3, Fig 4, Supplementary Fig 3-8,11,12, and supplementary Table 1-4,8,9.
Data under ERS700858, ERS700859, ERS700860, and ERS700861 are previous published data from other groups. These data are used for validation of our method. These data are associated with Fig 4.
Data under EGAS00001004373 are generated by our group. One part of these data are used for prognosis of immunotherapy on NSCLC patients using tumor-derived somatic mutations in cfDNA, as an application of our method; the other part of these data are used for the validation of our algorithm. These data are associated with Fig. 5, Supplementary Fig 6,9,10, and Supplementary Table 4.

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	We collected samples from 30 patients for analyses on prognosis markers for immunotherapy on NSCLC patients. This sample size was determined based on power analysis of Fisher's exact test on response rate between high-burden and low-burden groups. The response rate for the two groups were assumed based on Rizvi, Naiyer A., et al. Science 2016. With power of 0.85 and significance of 0.05, the sample size required per group is 13.
Data exclusions	We excluded 16 solid tumor biopsy sequencing data from the 30 patients, as they had extremely high duplication rate (>80%). Other samples were kept for analyses.
Replication	Due to the limited volumes of plasma DNA available, it was not feasible to perform sequencing assays more than once per sample. However, for 12 MBC patients and 6 CRPC patients, WES and WGS data was obtained to validate the tumor fraction for the same plasma sample, and WES data of two individual plasma samples and one tumor tissue biopsy sample were used to validate the mutations; for the 14 NSCLC patients, WES was performed for both the tumor tissue and the plasma to validate the mutations.
Randomization	The group allocation for the high-burden and low-burden groups can only be obtained when the analyses were done. Besides, our data were used for discovery and proof of concept. Randomization is not applicable.
Blinding	Blinding is not applicable.

Behavioural & social sciences study design

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

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*

Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

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

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access and import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data

Methods

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

Antibodies

Antibodies used — *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation — *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s) — *State the source of each cell line used.*

Authentication — *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination — *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See ICLAC register) — *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance — *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*

Specimen deposition — *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods — *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

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

Laboratory animals — *For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*

Wild animals — *Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*

Field-collected samples — *For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*

Ethics oversight — *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

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	KEYNOTE-001 is an open-label multicohort phase I study of 550 patients with treatment naive or previously treated locally advanced or metastatic NSCLC who were receiving pembrolizumab. Patients were randomly assigned a dose of either 2 mg or 10 mg per kilogram of body weight every 3 weeks or 10 mg per kilogram every 2 weeks. An additional cohort at the end were evaluated 2mg per kilogram of body weight every 3 weeks. KEYNOTE-010 is an open-label, international phase II/III study of 1034 patients with previously treated NSCLC and a PD-L1 expression ≥ 1% who were randomized to receive either pembrolizumab 2 mg or 10 mg per kilogram of body weight, or docetaxel 75 mg per meter squared every 3 weeks. Both studies have been previously reported in Edward BG et al. and Herbst RS et al. For both trials, eligible patients were ≥ 18 years, had an Eastern Cooperative Oncology Group (ECOG) performance status of 1 or less, and adequate organ function. Key exclusion criteria for both trials included a history of pneumonitis, prior immunosuppressive therapy, or active autoimmune disease. For KEYNOTE-001 trial, 53% participants are male, and the age range of all patients is 28-93; for KEYNOTE-010 trial, 61% participants are male, and the age range of all patients is 56-69. More detailed demographics of the patients from the KEYNOTE-001 study and KEYNOTE-010 study can be found in Edward BG et al. and Herbst RS et al. All patients from both trials used for this analysis were from UCLA. The bulk of the patients were enrolled in the KEYNOTE-001 study, and the UCLA patients from this trial have been previously published in Lisberg A et al. The demographics of the UCLA patients from the KEYNOTE-010 study are similar. For the 30 NSCLC patients in this analysis, 67% patients are male, and the age range of all patients is 32-86. We also purchased plasma samples and WBC samples of 3 healthy individuals from Biopartners, Inc. (Woodland Hills, CA). The donor cohort was comprised of two healthy males (23-year-old and 28-year-old) and one healthy female (52-year-old). The WBC samples and the plasma samples were obtained with appropriately consents of the healthy donors.
Recruitment	Potential patients were recruited as deemed clinically appropriate by the clinical investigators, per standard practices for clinical trials. Potential recruitment bias was limited by the specific inclusion and exclusion criteria in each protocol.
Ethics oversight	The specimens collected for the trials were collected in accordance with institutional biobanking protocols that were approved by UCLA Medical Institutional Review Board 2 (IRB# 12-001891, IRB# 11-003066, and IRB# 13-00394) and monitored by the Data Monitoring Committees. All patients provided written consent before any study-related procedures were performed.

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 ICMJEguidelines for publication of clinical research and a completedCONSORT checklist must be included with all submissions.

Clinical trial registration	ClinicalTrials.gov number NCT01295827 and NCT01905657
Study protocol	The full protocol for KEYNOTE-001 can be accessed at https://www.nejm.org/doi/suppl/10.1056/NEJMoa1501824/suppl_file/nejmoa1501824_protocol.pdf The full protocol for KEYNOTE-010 can be accessed at https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(15)01281-7/fulltext
Data collection	The participating sites were multiple academic medical centers in numerous countries. Patients were recruited in KEYNOTE-001 from May 2012 until February 2014, and data collection is ongoing. Patients were recruited between Aug 2013 and Feb 2015 in KEYNOTE-010, and data collection is ongoing. This analysis did not cover the entirety of KEYNOTE-001 and KEYNOTE-010, but only patients enrolled at UCLA. All patients in this analysis were treated and had data collected at UCLA.
Outcomes	In KEYNOTE-001 coprimary end points included objective response rate, median duration of response, progression-free survival, overall survival, as well as efficacy endpoints in patient with high PD-L1 expression levels. In KEYNOTE-010, The primary endpoints were overall survival and progression-free survival both in the total population and in patients with PD-L1 expression on at least 50% of tumor cells. Objective response and duration of response was measured using clinical assessment and radiographic assessment using the RECIST1.1 criteria. Progression-fee survival and overall survival were estimated using the non-parametric Kaplan-Meier method to estimate the curves in each treatment group. Death is always a confirmed progressive disease (PD) event and is censored in the analysis at the last disease assessment date. PD events are censored based upon 4 situations : 1. No PD and no death; new anticancer treatment is not initiated, 2. No PD and no death, but new anticancer treatment is initiated, 3. PD or death documented after ≤ 1 missed disease assessment, or 4.) PD or death documented after ≥ 2 missed disease assessments.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*

Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

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	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

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

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) — *Specify: functional, structural, diffusion, perfusion.*

Field strength — *Specify in Tesla*

Sequence & imaging parameters — *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition — *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software — *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization — *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template — *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal — *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring — *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings — *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested — *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) — *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction — *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity — *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis — *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis — *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*