nature research

Corresponding author(s): Marzieh haghighi
Shantanu Singh

Last updated by author(s): Aug 10, 2022

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 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	See "Supplementary A. Curated Datasets" section.
Data analysis	We used open source CellProfiler software (version 2.1 and 2.2, exact versions per data set are specified in Supplementary A) for extracting single cell features from images of each of the datasets and Cytominer (Cytominer v0.1.0 R 3.4.1) package for generating replicate level profiles . The code for analysis of data is public at: https://github.com/carpenterlab/2021_Haghighi_submitted

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 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

Preprocessed profiles that are augmented with gene and compound annotation are available through the Registry of Open Data on AWS on a public S3 bucket at no cost and no need for registration. Documentation on the folder structure, dataset details and instructions for accessing the data are available at http://broad.io/rosetta. Source datasets are described and referenced in Supplementary A.

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	Sample size was determined by practical limitations (how many cells fit in one sample well).
Data exclusions	No data were excluded. Data quality issues were noted in the paper.
Replication	For each provided dataset, each sample has at least 2 replicates. All replicate experiments performed are described in this paper, none were omitted.
Randomization	Treatments, including negative controls, were generally randomly distributed across well positions on the 384-well plates. We note two non-random patterns LINCS: All 6 doses of a treatment are in adjacent cells in the same row. This was due to constraints imposed by compound management robotics LUAD: All alleles of a gene are generally on the same plate. This was done intentionally so that comparing the wild-type and mutant forms of the gene were not subject to plate-to-plate variation.
Blinding	Image segmentation workflows were designed by experts who were blinded at the time to the identity of each sample; a small number of blinded samples were sub-selected to choose parameters to apply to all samples in a batch. All downstream steps were then performed identically for all samples in a data set.

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	CDRP utilized U-2OS cells from ATCC (HTB-96). LINCS utilized A549 cells from ATCC (CCL-185). TAORF utilized U-2 OS cells originally obtained from ATCC and propagated in the William Hahn lab. LUAD utilized A549 cells from ATCC and propagated in the Genetic Perturbation Platform at Broad Institute.
Authentication	The LUAD A549 cells were part of the Cancer Cell Line Encyclopedia project, which involved genetic/sequencing analysis. We are unaware of any other authentication done on other sets.
Mycoplasma contamination	Testing for mycoplasma was confirmed for the CDRP, CDRP-bio, and TAORF data sets. We lack the historical information to confirm whether the LUAD and LINCS sets were tested.
Commonly misidentified lines (See ICLAC register)	None are used (we used A549 and U2OS).