nature research

Corresponding author(s): Roman Schulte-Sasse, Annalisa Marsico

Last updated by author(s): Nov 15, 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 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

SNVs and DNA methylation were downloaded from TCGA using the GDC data transfer tool (manifest files are given in the github repository). Firehose-get was used to collect CNAs as output from GISTIC 2.0 from the latest Firehose repository. Gene expression data was collected from Wang et al. (2018), data record 3.

Furthermore, the following datasets were used to create positive and negative labels:
* Network of Cancer Genes (NCG) v6.0 (http://ncg.kcl.ac.uk/download_file.php?file=cancergenes_list.txt)
* DigSee database (http://210.107.182.61/digseeOld/)
* COSMIC Cancer Gene Census (CGC) v91 and COSMIC Mutations in Census Genes (https://cancer.sanger.ac.uk/cosmic/download)
* KEGG Cancer Pathways (https://www.gsea-msigdb.org/gsea/msigdb/cards/KEGG_PATHWAYS_IN_CANCER.html)
* OMIM disease genes (https://omim.org/downloads)

To validate EMOGI predictions, the following data sets were used:
* OncoKB database (https://www.oncokb.org/cancerGenes)
* ONGene database (http://ongene.bioinfo-minzhao.org/ongene_human.txt)
* Achilles Cancer Dependency Map CRISPR Genetic Dependencies (https://ndownloader.figshare.com/files/22629068)
* Predicted cancer genes from Bailey et al., 2018, Table S1 (https://www.cell.com/cms/10.1016/j.cell.2018.02.060/attachment/cf6b14b1-6af1-46c3-a2c2-91008c78e87f/mmc1.xlsx)
* DriverDB literature mining database for cancer driver genes (http://driverdb.tms.cmu.edu.tw/)
* Cancer initiation genes from CIGene (Liu et al., 2018; http://soft.bioinfo-minzhao.org/cigene/)
* Metastasis genes from Priestley et al., 2019, Table S5 (https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-019-1689-y/MediaObjects/41586_2019_1689_MOESM10_ESM.xlsx)
* Prognostic genes from Wee et al., 2018, Table S2 (https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-21691-5/MediaObjects/41598_2018_21691_MOESM3_ESM.xlsx)

Data analysis	Batch effects were corrected for using ComBat (sva R package), DNA methylation data (TCGA level 3) was preprocessed using custom python code. SNV data (also TCGA level 3) was further corrected for known hyper-mutated samples (listed in syn1729383 on synapse) and otherwise preprocessed using custom code along with CNA information. The GCN python package was used as base for a custom GCN implementation. Custom code was used for model training and data analysis.

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

The datasets analysed during the study are available via TCGA data portal (https://portal.gdc.cancer.gov/), Firehose (https://gdac.broadinstitute.org/), Figshare (https://doi.org/10.6084/m9.figshare.5330593). The several PPI networks used are publically available, e.g. the CPDB PPI network (http://cpdb.molgen.mpg.de/). A complete data availability statement is presented in the manuscript.

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	All available samples were downloaded from TCGA for the selected cancer types/studies (numbers are given in table S1). Because we average across all samples, the sample size is sufficient to build a reasonable estimate across cohorts. Training, test and validation sets are chosen using label stratification. 75% of the genes are used for training and 25% for testing. Of the 75% training genes, 20% were used for validation. This is a standard procedure and similar numbers are used throughout the literature.
Data exclusions	Hyper-mutated samples were removed as listed in synapse (syn1729383) which is a standard procedure and done in Leiserson et al., 2014 and similar studies.
Replication	Cross-validation of the models was used to assess robustness (technical replication). To further ensure reproducibility of the results, independent cancer gene sets were collected and compared against and models were trained on 6 different PPI networks.
Randomization	10-fold cross validation and perturbation experiments were used to assess model robustness.
Blinding	No blinding was used in this study.

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