nature research

Corresponding author(s): Neville Sanjana and David Knowles

Last updated by author(s): May 10, 2023

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	A full data collection description is provided in the method section: Sequencing reads were trimmed using a custom python (v.2.7.10) script (available upon request). gRNA read counts were generated using FASTX-Toolkit (v0.0.14). gRNA and target RNA feature extraction: RNAfold and RNAplfold ViennaRNA (v2.4.10) RNAhybrid (v2.1.2) gRNA prediction from alternative guide RNA prediction algorithm were collected here: http://deepcas13.weililab.org (No version ID available) https://www.rnatargeting.org (No version ID available) https://cas13design.nygenome.org (v1.1, re-trained on data generated in this study where indicated in the text)
Data analysis	The method section describes in detail data processing and analysis: Pooled screen data processing: R v3.6.0 RobustRankAggreg v1.1 (R package) SVA v3.34.0 (R package) Deep Learning: We use TensorFlow's official Docker image for version 2.11.0 with GPU support: https://hub.docker.com/layers/tensorflow/tensorflow/2.11.0-gpu/images/ sha256-67f1a7b[...]1c0cd311655be7477f2bc1b6f27e014b9a57231bd55b3?context=explore Additional python packages not included in Docker image: bbiopython (version 1.80) pandas (version 1.5.2) matplotlib (version 3.6.1)

seaborn (version 0.12.1)
shap (version 0.41.0)
statsmodels (version 0.13.5)
tensorflow-probability (version 0.19.0)
statsmodels (version 0.12.2)
sklearn
We have made all code essential to this study available on https://github.com/daklab/tiger as noted in the Code Availability Statement.

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

All data generated in this study has been deposited at NCBI Gene Expression Omnibus (GEO) with the accession number GSE232228. Raw flow cytometry screen data from Wessels, Méndez-Mancilla et al. (ref. 7) is available under the accession number GSE142675.

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	A full description of Cas13d guide RNA library size and composition is provided in the method section. All screens have been conducted in at least three independent replicate experiments. Detailed statistics for included guide numbers and guide classes can be found in Supplementary Data 2, 5 and 10. The final predictive guide model (TIGERcombined) was constructed using features for 93,145 guides RNAs (Types: PM, SM, DM, TM, RDM, RTM) (255 filtered gRNAs of the same gRNA types removed) that match to coding sequences of the 16 target genes presented in Supplementary Data 1 and 2.
Data exclusions	The initial HEK293FT off-target screen was conducted in three replicates. In total, we removed 333 targeting gRNAs due to filtering steps described in the method section. In addition, non-reproducible technical outliers were masked by flagging individual values with high variance within replicate samples of each timepoint (D0, D15, D30). Masking did not lead to gRNA removal.
Replication	All screens have been conducted in at least three replicate experiments. We confirm that all screen have been conducted independently. Replicate screens were highly correlated.
Randomization	We consider three cross-validation methods: one at the gene level, one at the target level and one for technical holdouts. For gene holdouts, we create sixteen folds where each fold contains all (guide, target) tuples specific to that gene. We feel this method is generally the most challenging method to succeed on as it requires and promotes transcriptome-wide generalization. The second method randomly folds target sites into ten, non-overlapping folds. Holding out target site places all perfectly matched and mismatched gRNAs for a target site into the holdout set, ensuring that a target sequence never appears both in training and validation sets. Without this restriction (i.e. randomly folding at the gRNA level), we found that more sophisticated model architectures, specifically those with recurrent units, that perform extremely well on gRNA holdouts (Pearson r = 0.88) failed to generalize to held out genes (Pearson r = 0.36). For the technical holdout, we hold out gRNAs from the flow cytometry of cell surface proteins from three genes. Thus, our technical holdout is a multi-gene holdout. For the technical holdout, we used the log2-transformed gRNA depletion in the fluorophore-low bin divided by the fluorophore-high bin.
Blinding	Not applicable. All experiments have been processed and analyzed in an unbiased way. Analysis did not require blinding.

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)	HEK293FT cells (also denoted to as HEK293 cells) were acquired from Thermo Fisher (R70007). HAP1 cells were acquired from Horizon Discovery.
Authentication	The cell lines were not authenticated in the lab after purchase from the vendors.
Mycoplasma contamination	All cell lines were tested as mycoplasma-free using Lonza MycoAlert (#LT07-518).
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.