nature research

Corresponding author(s):	Yi Xing
Last updated by author(s):	Jan 23, 2019

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	Publicly available data were downloaded from ENCODE/Roadmap data portal or from GEO using corresponding GEO accession IDs as stated in the manuscript. Data download was performed by the unix program wget(v1.12).
Data analysis	The DARTS software code, predictive features and trained model parameters are freely available in GitHub as stated in the manuscript. DARTS was developed and tested in Python(v2.7) and R(v3.4.3). DARTS deep learning model was implemented in Keras(v2.0.6). RNA-seq raw data was processed by STAR(v2.5.2a), Kallisto(v0.43.0), rMATS(v4.0.1), MISO(v0.5.3). Simulated RNA-seq data was generated by Flux-simulator(v1.2.1). RASL-seq raw data was processed by Blat(v36x1). Predictive features by prediction tools were generated by NuPoP(1.24.0), Maxent(20-Apr-2004), RNAfold(2.2.10).

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

High-throughput sequencing data of PC3E and GS689 cell lines that support the findings in this study were deposited into GEO with the accession ID GSE112037. Publicly available data were downloaded from ENCODE/Roadmap data portal or from GEO using corresponding GEO accession IDs as stated in the manuscript. Computer programs and software are available in GitHub https://github.com/Xinglab/DARTS.

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	No sample size calculation was performed. Publicly available data in ENCODE/Roadmap download portal were used (as of May 2017).
Data exclusions	No data were excluded.
Replication	Three biological replicates of RNA-seq experiments were performed on PC3E and GS689 cell lines. No replication failed.
Randomization	This is not relevant to our study, because our study does not involve the assignment of test subjects or treatments.
Blinding	This is not relevant to our study, because our study does not involve the assignment of test subjects or treatments.

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	PC3E and GS689 cell lines were derived from the parental PC3 cell line (ATCC CRL-1435). PC3E was derived from PC3 cells by fluorescence-activated cell sorting for E-cadherin-positive cells, and GS689 was recovered from a secondary metastatic liver tumor after intravenous injection of PC3 cells into mouse (Mol Cancer Res. 2015 13(2):305-18).
Authentication	Authentication for all cell lines used in this study was performed by Laragen, Inc using the Promega powerplex16 System recommended by American Type Culture Collection. The STR alleles were searched either on ATTC or DSMZ databases depending on availability of the cell lines in the databases.
Mycoplasma contamination	Cell lines were not tested for Mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	Not applicable.