nature research

Corresponding author(s): Hyongbum Henry Kim

Last updated by author(s): Dec 4, 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	Illumina HiSeq and MiniSeq Sequencing software were used for the collection of deep-sequencing data.
Data analysis	Custom Python scripts were used for random-sequence generation, indel-frequency analysis, and for the implementation of DeepxCas9 and DeepSpCas9-NG. Statistical significance was determined by using PASW Statistics (version 17.0, IBM) and Microsoft Excel (version 16.0, Microsoft Corporation).

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

The authors declare that all data supporting the results in this study are available within the paper and its Supplementary Information. The deep-sequencing data from this study is available at the NCBI Sequence Read Archive under accession number SRP158724.

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 statistical methods were used to predetermine sample size. Sample sizes were chosen after deep sequencing, depending on the read number and quality. All sample sizes were sufficient for the model training and for the following statistical tests, as evidenced by the high Spearman correlation coefficient measured for the cross-validation and by the statistical values for the Student's t-test, One-way ANOVA and post-hoc tests.
Data exclusions	To increase the accuracy of the analysis, we excluded data according to the filtering conditions for the indel-frequency analysis. The filtering conditions are described in Methods. The criteria were established after deep sequencing, depending on the read number.
Replication	High-throughput experiments were replicated using different MOI of the Cas9-expressing lentivirus. Results are provided in Supplementary Datasets 1 and 2, and in Supplementary Figs. 6 and 11.
Randomization	We selected endogenous target sites by random sampling, for the fair evaluation of the performance of DeepxCas9 and DeepSpCas9-NG.
Blinding	The investigators were not blinded to group allocation. This study does not involve animals or human research participants.

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)	HEK293T, from American Type Culture Collection (ATCC). HCT116, from Korean Cell Line Bank (KCLB).
Authentication	Cells were not authenticated.
Mycoplasma contamination	The cells were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.