nature research

Corresponding author(s): Hyongbum Henry Kim

Last updated by author(s): Aug 11, 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	Targeted deep sequencing data were collected using HiSeq2500, HiSeq 4000, and MiniSeq (Illumina).
Data analysis	GraphPad Prism 8, PASW Statistics (version 18.0, IBM), Microsoft Excel (version, 16.0, Microsoft Corporation), XGBoost Python package (version 0.90), scikit-learn (version 0.19.1), and TensorFlow were used. Source codes for DeepPE and custom python script used for the prime editing efficiency calculation are provided as Supplementary codes 1 and 2 and also available at https://github.com/hkimlab-PE/PE_SupplementaryCode.

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 deep sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number PRJNA624815.

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 following statistical test.
Data exclusions	To increase the accuracy of the analysis for PE efficiency, deep sequencing read counts were below 200 or the background PE frequencies were above 5% were excluded.
Replication	We replicated experiments as described in the manuscript. We performed high-throughput evaluation in duplicate by two independent researchers. We evaluated PE2 efficiencies at endogenous sites of HEK293T cells in sextuplicate and those of HCT116 and MDA-MB-231 cells in quadruplicate. All replcates were performed successfully and showed strong correlation between replicates.
Randomization	We selected target sequences for the development of DeepPE and other conventional machine learning based models by stratified random sampling.
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 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)	The source of the cell line, HEK293T, is American Type Culture Collection (ATCC). The source of the cell line, HCT116 and MDA-MB-231, are Korean Cell Line Bank (KCLB).
Authentication	Not been authenticated.
Mycoplasma contamination	Not been tested.
Commonly misidentified lines (See ICLAC register)	HEK293T, HCT116, and MDA-MB-231 are not listed in the ICLAC.