

Reporting Summary

Nature Portfolio 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 Portfolio 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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give <i>P</i> values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	Next-generation sequencing data was collected and demultiplexed by Illumina NovaSeq, Huada MGI DNBSEQ and Salus EVO Platform
Data analysis	All custom code used for analysis are available at github: ContactSeek (version 1.0.0; https://doi.org/10.5281/zenodo.20487746; https://github.com/menghaowei/ContactSeek) dl-profiling toolbox (version 1.0.0; https://doi.org/10.5281/zenodo.20487721; https://github.com/menghaowei/dipro) Cutadapt (version 5.1) BWA (version 0.7.17) samtools (version 1.15.1) Picard (version 2.27.0) samclip (version 0.4.0) HISAT-3N (version 0.0.3) STAR (version 2.7.11b) R (version 4.4.1) AlphaFold3 (version 3.0.0) OpenMM (version 8.2) FlowJo (version 10.8.1)

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 Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Datasets generated in this study have been deposited in public repositories. Both the AlphaFold3 prediction outputs (zenodo.17197986 and zenodo.19555204) and the targeted amplicon sequencing raw count tables (zenodo.17243125) are available on Zenodo. The raw dl-profiling sequencing data have been deposited in the Genome Sequence Archive (GSA) in National Genomics Data Center (NGDC) under accession code HRA013626 and HRA013767 (BioProject PRJCA046878). RNA-seq data for RNA off-target assessments of A3A and TadA8e were downloaded from the SRA database under accession codes PRJNA938578 and PRJNA811343, respectively. Targeted amplicon sequencing data for A3A variant screening were obtained from the SRA database under accession code PRJNA938162. The ChIP-seq and ATAC-seq data related to chromatin state analysis for off-target sites were downloaded from the ENCODE (ENCFF993NDR, ENCFF480MMN, ENCFF400CMC, ENCFF449FCR, ENCFF567FDM) and GEO (GSM1816075, GSM317585, GSM1816068, GSM1816079, GSM317589, GSM8091609, GSM3762822) databases. These public datasets are listed in the Supplementary Table 6. In this study, the human reference genome assembly GRCh38 (hg38) was used as the reference genome.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	Not relevant
Reporting on race, ethnicity, or other socially relevant groupings	Not relevant
Population characteristics	Not relevant
Recruitment	Not relevant
Ethics oversight	Not relevant

Note that full information on the approval of the study protocol must also be provided 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](https://www.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 method was used to predetermine sample size. For the dl-profiling method, we evaluated its reproducibility and robustness across 26 cell-line samples and 5 HPSC samples. Technical replicates showed strong concordance across all samples (Supplementary Figs. 2-7), demonstrating that the assay yields highly consistent results at the cellular level. For the Detect-seq method, we have also rigorously benchmarked its robustness in our previous studies (Refs. 27, 28, and 58). On this basis, the current results from dl-profiling and Detect-seq are considered sufficient to support the conclusions of this study. For all other experiments in this study, including variant screening and characterization of the best-performing variants, assays were performed at the cell-line level using transfection and sample-preparation protocols consistent with established practice in the base editing field, with no special settings applied. On-target and off-target editing rates measured by deep sequencing in such systems have been shown to be highly reproducible across replicates (e.g., Refs. 6, 7, 9 and 10). Accordingly, a sample size of $n \geq 2$ biological replicates was used, which is considered sufficient to support the conclusions of this study.
Data exclusions	No data were excluded from the analyses.
Replication	Findings were reproduced in independent biological replicates performed at intervals ranging from weeks to months. All experiments were repeated at least once, and all independent replicates yielded consistent results supporting the conclusions presented.
Randomization	Randomization was not applied, as is typical for cell culture experiments in the genome editing field (e.g., Refs. 6, 7, 9 and 10). Editing outcomes were quantified by high-throughput sequencing and analyzed with previously validated bioinformatic pipelines using default settings, minimizing the potential for selection bias.

Blinding

Blinding was not applied, as is typical for cell culture experiments in the genome editing field (e.g., Refs. 6, 7, 9 and 10). All readouts were quantitative (deep sequencing-based) and analyzed with previously validated bioinformatic pipelines using default settings, minimizing the potential for experimenter bias in outcome assessment.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	HEK293T cells (source SUNNCELL, SNL-015), U2OS cells (source: SUNNCELL SNL-054), and HCT116 cells (source: SUNNCELL SNL-077). Frozen healthy CD34+ cells were purchased from MileCell Bio (China)
Authentication	All cell lines are from authenticated manufacturers.
Mycoplasma contamination	Cells tested negative for mycoplasma as detailed in the Methods.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Plants

Seed stocks	Not relevant
Novel plant genotypes	Not relevant
Authentication	Not relevant

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells were seeded into 24-well or 6-well plates and transfected with plasmids at 70% confluence using Lipofectamine 2000, following the manufacturer's instructions. Two days after transfection (for RNA-seq samples), three days after transfection
--------------------	---

(for CBE editing), or five days after transfection (for ABE editing), cells were washed with 1× PBS and dissociated into single-cell suspensions using 0.25% Trypsin-EDTA (Gibco). Cells were then washed twice with PBS and resuspended at a concentration of approximately 1E7 cells per mL. Before flow cytometry, cells were filtered through a 35 µm cell strainer cap (Corning).

Instrument

BD Aria SORP or BD Aria Fusion

Software

BD FACSDiva 9.0.1 and FlowJo 10.8.1

Cell population abundance

Cells with the desired EGFP signals were sorted. For dl-profiling samples subjected to FACS, approximately 800,000 cells were collected. For RNA-seq samples treated with Cas12a-CBE, approximately 200,000 cells were collected. For Cas12a-CBE editing in U2OS cells, 1,000–2,000 cells were collected. For Cas12a-CBE editing in HCT116 cells, approximately 10,000 cells were collected. For each sample, data from 10,000 to 40,000 total events were recorded.

Gating strategy

Cells were sequentially gated to exclude debris and doublets based on FSC and SSC. For dl-profiling samples subjected to FACS, cells with the top 20% EGFP (FITC-A) intensity were collected. For RNA-seq samples treated with Cas12a-CBE, cells with the top 15% EGFP (FITC-A) intensity were collected. For Cas12a-CBE editing in U2OS and HCT116 cells, EGFP (FITC-A)-positive cells were collected.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.