nature portfolio

Corresponding author(s): Xiaolong Wang, Jianghuai Liu and Lina Wu

Last updated by author(s): Apr 30, 2026

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 (*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 NovaSeq 6000 (2x 150PE), pacbio
Data analysis	Targeted amplicon sequencing data were first processed using AdapterRemoval (version 2.2.2) for barcode recognition and sample demultiplexing, resulting in standard paired-end sequencing files (pair1 and pair2). Subsequent genome editing analysis for each sample was performed using CRISPResso2. For samples containing expected homology-directed repair (HDR) insertions, the reference amplicon sequence (amplicon_seq), spacer sequence, and the expected HDR sequence were provided to enable HDR analysis mode in CRISPResso2 (activated via the -e parameter). The analysis was conducted using the default quantification window (--quantification_window_size 4) and plot window (--plot_window_size 30). Final outputs included HDR efficiency, indel frequency, editing profiles, and visualization plots. Targeted amplicon sequencing data for off-target effect evaluation were preprocessed in the same manner as on-target samples. Subsequent analysis of indel frequencies was performed using CRISPResso2. For each site, the reference amplicon sequence (amplicon_seq) and corresponding spacer sequence were provided. A cleavage offset of –3 (--cleavage_offset -3) was specified to simulate the Cas nuclease cut site, and editing quantification was carried out within a 4 bp window upstream and downstream of the cleavage site (--quantification_window_size 4). The final outputs included indel frequency statistics at each site, sequence alignment results, and visualization plots, enabling assessment of editing activity at potential off-target loci. GUIDE-seq sequencing data were first processed with Cutadapt for adapter trimming and quality filtering, followed by PCR duplicate removal using unique molecular identifier (UMI) tags. Clean reads were aligned to the human reference genome using BWA, and potential off-target sites were identified through GUIDE-seq–specific analysis scripts. All cleavage sites were subsequently annotated for genomic features and cancer gene associations using ANNOVAR. The analysis yielded a genome-wide map of double-strand break (DSB) sites and on/off-target locations, enabling comprehensive assessment of off-target activity. UDiTaS sequencing reads were first subjected to quality control and removal of adapter sequences. Donor sequences were then searched within the reads and the downstream genomic flanking sequences were extracted. Specifically, local alignment was performed using edlib,

allowing up to four edit distances including mismatches or indels to detect the donor sequence or its reverse complement. For reads containing donor sequences, the downstream genomic region adjacent to the donor was extracted, and only flanking sequences of at least 30 bp were retained for further analysis. The extracted flanking sequences were subsequently aligned to the human reference genome GRCh38 using BWA-MEM. For successfully mapped reads, the 5' end position of each read was defined as the genomic integration site. To reduce the impact of sequencing errors or minor positional shifts, integration sites located within ±5 bp were clustered and considered as a single integration event. To increase confidence in the identified sites, only integration sites detected in both biological replicates were retained, and a read frequency threshold of 0.015% was applied. The on-target site at chr14:22766901 was excluded from the final analysis.
The pacbio resulting CCS reads were first demultiplexed according to the sequencing barcodes and PCR primer sequences, and then aligned to a reference sequence containing the target genomic region and the donor DNA. Sequence alignment was performed using minimap2, followed by sorting and indexing of the alignment files with samtools. Only reads with primary alignment and mapping quality ≥ 20 were retained for downstream analysis. Indels events were extracted from the CIGAR strings of aligned reads and their positions in the reference sequence were recorded. For each nucleotide position across the integration region, the sequencing coverage, the number of reads containing indels, and the corresponding indel frequency (indel reads divided by total coverage) were calculated. A per base indel profile across the integration region was then generated to assess the sequence integrity and structural stability of the donor DNA and its junctions with the genomic DNA.

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

The GUIDE-seq,pacbio and deep-sequence data have been deposited in the NCBI SRA under Bioproject numbers PRJNA1307969 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1307969]. Source data are provided with this paper.

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	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

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

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes are indicated in the figure legends. No statistical methods were used to predetermine sample sizes. All experiments were performed under rigorously controlled conditions with validated reagents to minimize non-specific interference and experimental artifacts. Unbiased deep sequencing was employed throughout the study to ensure robust and high-resolution readouts. With the reported sample sizes, the observed trends were consistent and reproducible across replicates. Importantly, the sample sizes used (typically n = 3) were sufficient to support the statistical significance of key findings.GUIDE-seq analyses were conducted with one replicate per group.ddPCR analyses were conducted with two replicates per group.
Data exclusions	No data has been excluded from the analyses.
Replication	As indicated in the figure legends, all presented data have been validated through biological replicates and independent experiments. Results are shown as mean±SD from replicate measurements. The findings are highly reproducible.
Randomization	For each experiment in this study, all cells originated from the same initial culture. They were distributed into multiple plates and randomly allocated to different treatment groups (e.g., transfection with specific plasmids). Apart from the designated treatments, all cells were maintained and handled under identical conditions.
Blinding	Blinding was not implemented in this study, due to the exploratory nature of the relevant experiments. In most experiments, samples from different groups were processed and analyzed in parallel within the same experimental run. The measurements—such as fluorescence quantification and deep sequencing—are inherently objective. Furthermore, experimental replicates were conducted to ensure the reliability of results and to rule out potential artifacts.

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

Antibodies

Antibodies used	MYC tag Polyclonal antibody (Proteintech (16286-1-AP), diluted 1:2000), GAPDH Polyclonal antibody (Proteintech (10494-1-AP), diluted 1:5000) and Goat Anti-Rabbit Recombinant Secondary Antibody (H+L)(Proteintech (SA00001-2), diluted 1:5000)
Validation	The MYC tag Polyclonal antibody and Goat Anti-Rabbit Recombinant Secondary Antibody (H+L) have been characterized by the vendor (https://www.ptgcn.com/). Our results in Supplementary Fig. 5b, e also showed bands at the expected sizes.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HEK293T (ATCC CRL-3216), HCT116 (ATCC CCL-247), JeKo-1 (ATCC CRL-3006), JurKat (ATCC TIB-152), Huh-7 (Cellosaurus CVCL_0336), hTERT RPE-1 (ATCC, CRL-4000)
Authentication	HEK293T and HCT116 cells were not authenticated. JeKo-1, Jurkat, REP-1 and Huh-7 cell lines were authenticated by the supplier using STR profiling.
Mycoplasma contamination	All cell lines were tested negative of mycoplasma contaminations.
Commonly misidentified lines (See ICLAC register)	None of the cell lines used are listed in the ICLAC database.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

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	Transfected cells were harvested by trypsinization to obtain single-cell suspensions, which were then analyzed by flow cytometry.
Instrument	BD LSRFortessa X20
Software	FACSDiva for collections, and FlowJo for analyses
Cell population abundance	For analyses of reporter plasmid, a total of around 10,000 cells were analyzed.
Gating strategy	In cells transfected with the reporter plasmid for assessing the recombination efficiency of IS621 and its variants, the main cell populations were first identified and gated based on their distribution in the FSC/SSC plot. Transfected cells were then further gated according to GFP^+ (reporter plasmid) and $mCherry^+$ (IS621 or its variants) signals, using untreated cells as the reference ($EGFP^-$) to define the gate for recombination-responsive $EGFP^+$ cells.

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