nature research

Corresponding author(s): Jay Shendure, Lea M. Starita

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used except for Illumina RTA basecalling.
Data analysis	Common, freely available DNA sequencing data analysis software was used to analyze data, as described in Methods: bcl2fastq v2.16, Python 2.7.3, SeqPrep (available at https://github.com/jstjohn/SeqPrep), fastqc v0.11.3, EMBOSS v6.4.0, R v3.1.3, RStudio v1.0.153. Custom scripts were written in Python and R to analyze data, and are available at: https://github.com/shendurelab/saturationGenomeEditing_pipeline.

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 upon request. 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

Function scores are freely available for all nonprofit uses (see https://sge.gs.washington.edu/BRCA1/), as well as by nonexclusive license under reasonable terms to commercial entities that have committed to open sharing of BRCA1 sequence variants. Sequencing data is available at GEO under accession GSE117159. Custom scripts for analyzing sequencing data were written in Python and R. All code is available at: https://github.com/shendurelab/saturationGenomeEditing_pipeline.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	No explicit calculations were performed to determine sample size. Rather, we aimed to test every possible nucleotide in each of the 13 BRCA1 exons that encode for the RING and BRCT domains. Therefore, in each transfection, the number of SNVs tested was roughly 3-fold the number of nucleotides in the exon tested (including some adjacent intron sequence), corresponding to all possible SNVs. Our achieved sampling is shown in Extended Data Fig. 3c. To effectively sample each SNV in each experiment, we transfected cell populations large enough to maximize the chance every library SNV was introduced to the genome hundreds of times each (approximately 20,000,000 cells per experiment).
Data exclusions	Exclusion criteria were not pre-established. Exclusions were performed prior to analyzing the accuracy of the data at predicting clinical variant interpretations. Conclusions remain the same with or without exclusions, as shown in Extended Data Fig. 6e-g. Approximately 3.5% of SNV measurements were excluded. In short, data were excluded according to uniformly applied rules to ensure scores were reflective of experimental selection and not poor sampling and/or artifacts resulting from the genome editing process (Extended Data Fig. 6a). More specifically, to rule out artificial sampling from sequencing error, SNVs not adequately sampled on Day 5 (over 1 in 100,000 reads) were excluded. Certain SNVs located near Cas9 target sites made the enzyme more likely to re-cut the genome after editing, thereby artificially lowering the SNVs abundance (Extended Data Fig. 6b,c). Additionally, SNVs that occurred near HDR marker mutations and due to this, caused different amino acid sequence changes or had greater potential to disrupt splicing were also excluded. RNA data for exon 18 was excluded due to a lack of reproducibility across replicates (Extended Data Fig. 4), and exon 22 WT data was excluded due to primer cross-reactivity that led to an uncharacteristically high correlation between plasmid library measurements and Day 5 measurements (Extended Data Fig. 4a).
Replication	All SGE experiments were performed at least two times to assess reproducibility, which is described in Extended Data Fig. 4. A small fraction of SNVs scored discordantly between replicates were removed from the data set (n=14 SNVs, Extended Data Fig. 6a). Two replicates were obtained in both WT HAP1 cells, and HAP1-Lig4KO cells. Results between cell lines also showed a high correlation (Extended Data Figs. 3i and 10c).
Randomization	Variants were tested in multiplex, with hundreds of variants related by physical proximity in the genome per sample.
Blinding	Function scores for each variant were determined without knowledge of known pathogenicity status, which was only assessed after.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

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)	HAP1 cells were obtained from Haplogen (now Horizon Discovery).
Authentication	Cells lines were authenticated by karyotype.
Mycoplasma contamination	Cell lines were not tested for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

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	Live cell, DNA content staining with Hoescht dye
Instrument	BD FACS Aria II
Software	FlowJo10
Cell population abundance	The haploid, 1N population was 29%
Gating strategy	The 1N population was identified by staining for DNA content.

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