nature research

Corresponding author(s): Jinsong Li, Yong Chen

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	Images were collected using DP-BSW (Version 03.02) of OLYMPUS. Western blot signals were detected using an Image Lab 4.1 on the ChemiDoc MP of BIO-RAD.
Data analysis	SnapGene (Version 2.3.3) was used to analyse the sanger sequencing results. GraphPad Prism 7 and Microsoft Office Excel (office 2016) were used to perform statistical analyses. Flow Jo. 7.6.1 was used to perform FACS data. Image J was used to handle the images. GuideScan (http://www.guidescan.com/) was used to design the sgRNA library. Homology structural models were generated from I-TASSER30 (https://zhanglab.ccmb.med.umich.edu/I-TASSER/), PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) and Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2) services and drawed by PyMOL 2.0.

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

The deep sequencing and whole-exome sequencing data from this study have been deposited in the Gene Expression Omnibus (GEO) under accession codes GSE115015 and GSE115017. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

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 statistical method was used to predetermine sample size. For the mouse experiments, 3 animals per condition are required to show the phenotype. For most experiments, we exceeded this sample size to confirm our results. Results obtained were highly significant and consistent and did not require larger animal experimental groups.
Data exclusions	No data were excluded from the analysis.
Replication	All replication attempts were successful.
Randomization	In vitro experiments were not randomized. In vivo experiments were randomized.
Blinding	The investigators were blinded to group allocation during in vivo experiments and outcome assessment.

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

Antibodies

Antibodies used	The following antibodies were used for iimmunoblotting: anti-Flag at 1:1000 (mouse monoclonal clone M2, Sigma, Cat#F1804, Lot#SLBQ6349V), anti-HA at 1:1000 (rabbit monoclonal, CST, Cat#3724, Lot#8), anti-actin at 1:10000 (rabbit monoclonal, Abclonal, Cat#AC026, Lot#0201920101), anti-PCNA actin at 1:1000 (rabbit polyclonal, Abclonal, Cat#A0264, Lot#9100026001). The following secondary antibodies were used: Goat anti-Mouse IgG-HRP conjugate (1:5000, Cat#115-035-003) from Jackson. Goat anti-Rabbit IgG-HRP conjugate (1:5000, Cat#111-035-003) from Jackson.
Validation	Validations are based on the datasheets from the manufacturers. We also tested the validation by using positive and negative cells.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Haploid ESCs (408) was established by ourself. E14 was obtained from Peter Mombaerts' lab at the Rockefeller University. HEK293T cells were obtained from Cell Bank at SIBCB.
Authentication	408 has been established and used in our previous studies (Cell Stem Cell, 2015, 17, 221). E14 has been tested in my previous study (Nature, 2004, 428, 393). HEK293T cells have been used in our previous study (Cell Stem Cell, 2015, 17, 221).
Mycoplasma contamination	Cell lines were tested to be free of mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Oct4-EGFP males: C57BL/6, age 9-14 weeks; B6D2F1 (C57BL/6 X DBA2) females: age 7-9 weeks; C57BL/6 males: age 9-14 weeks; ICR females: age 8-10 weeks;
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field-collected samples.

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	Preparation of single cell suspension from 4B2N1: 4B2N1 were trypsinized with 0.05% Trypsin-EDTA in 37 °C for 5min, and then incubated with 15 ug/ml Hoechst 33342 in 37 °C water bath for 10 min.
Instrument	BD FACS Aria II
Software	FlowJo 7.6.1 was used to analyze FACS data.
Cell population abundance	The single cell suspension from 4B2N1 was sorted in BD FACS Aria II with purity of priority. The double positive cells (RFP positive and the first peak of DAPI) were enriched only.
Gating strategy	Step one, gating living cells used FSC-A vs SSC-A and SSC-W vs SSC-H; Step two, gating PerCP-A positive cells; Step three, gating sorting 1N peak of DAPI.

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