nature research

Corresponding author(s): Volker Hovestadt, Bradley E. Bernstein

Last updated by author(s): 3/29/2021

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	Data was generated using the Illumunia NextSeq500 sequencing instrument according to the manufacturer's guidelines. Fastq files were generated using bcl2fastq version 2.15.0.
Data analysis	Raw sequencing data was processed using Cutadapt version 2.7 and TrimGalore version 0.6.5. Sequencing data was then aligned to the human and/or mouse genome using methylCtools version 1.0.0 and bwa mem version 0.7.17. Further analysis were performed using R version 3.6.2. Sorting of single cells based on cell surface markers was analyzed with FlowJo version 10.6.0. Illumina SNP array data was processed using GenomeStudio version 2.0.4.

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

All data generated as part of this study is available as processed DNA methylation calls and as raw sequencing data from GEO/SRA under the accession number GSE149954. Furthermore, a number of public datasets were used and an overview is provided in Supplementary Table 5.

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	Not applicable, sample size as described in text and methods.
Data exclusions	No data was excluded in this study
Replication	Initial experiments in K562 cells (from 10ng purified DNA) were replicated and showed high correlation between independent experiments. Other cell lines were not technically replicated. However, profiles of a cell line from different starting input amounts showed high correlation.
Randomization	Not applicable, no randomization was performed. All samples were treated the same in library preparation.
Blinding	Not applicable, no blinding was performed. Different treatments and controls were profiled together in the same batch.

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)	All cell lines described in this study were obtained from ATCC (K562, Yac1, Kasumi1, OCI-AML3, HL-60, GM12878).
Authentication	Copy-number variations identified from DNA methylation data matched expected profiles described in literature.
Mycoplasma contamination	All cell lines used in this study tested negative for Mycoplasma contamination using PCR.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics	A deidentified sample of human bone marrow was selected for this study. Unknown age, gender, genotypic information, past and current diagnosis and treatments.
Recruitment	Bone marrow donors consented to an excess sample banking and sequencing protocol that covered all study procedures.
Ethics oversight	Institutional Review Board (IRB) of the Dana-Farber/Harvard Cancer Consortium

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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 resuspended in PBS and 2% FBS and stained on ice for 20 minutes. Cells were then washed with PBS and 2% FBS.
Instrument	Sony SH800
Software	FlowJo version 10.6.0
Cell population abundance	Bone marrow cells were sorted based on CD34+, CD3+, and CD14+ staining. Manual compensation was conducted using single loaded antibodies on flow beads. An unstained sample of cells and a sample containing PI staining were used as additional controls. Voltages of channels remained constant throughout the experiment. Single viable cells were initially filtered. CD34+ cells were CD34 positive and CD14 negative. Similarly, CD3+ cells were CD3 positive and CD14 negative, while CD14+ cells were CD14 positive and CD3 negative.
Gating strategy	Gating strategy is provided in Supplementary Figure 8A. In short, filtering was based on SSC-A vs FSC-A (positive: FSC-A>200K), FSC-W (positive : FSC-W <300) vs FSC-H for singlets, PI vs FSC-A (positive PI < 2x10e2) for viable cells.

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