nature research

Corresponding author(s): Dan Landau

Last updated by author(s): Mar 14, 2019

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.

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	No software was used for data collection.
Data analysis	Bismark (v.0.14.5); bowtie2-2.2.8; BWA v0.7.17; Picard; Macs2 (v2.0.10); STAR(version 2.5.2a); ZINBWaVE(v. 1.0.0); edgeR (v. 3.20.1); GSEA software and Molecular Signature Database (MSigDB) (http://www.broad.mit.edu/gsea/); BEDTools v2.25.0; MEME-ChIP; IQ-TREE v1.5.3; FigTreev1.4.3; Python 2.7.13; R version 3.4.2.

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. 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

MscRRBS and single-cell Smart-seq2 datasets have been deposited to the NCBI Gene Expression Omnibus (GEO) under accession number GSE109085. ChIP-seq datasets have been deposited to the NCBI GEO under accession number GSE119103. Additional supplementary data is available upon request.

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	We sequenced 2,652 single cells in total from 6 and 12 B cell healthy donors and CLL patients, respectively, enabling us to do statistics both at the single cell and sample level, giving us enough statistical power to detect differences in all the analyses reported in this study (e.g., epimutation rates difference, concordance odds ratio analysis, four-gamete analysis). In addition, PAC learning analysis showed that ~400K CpGs enable trees with up to 350 leaves, allowing for information loss due to random sampling of two alleles. The dataset has >80% power to detect significant ($p < 0.05$), epigenetically derived subpopulations.
Data exclusions	No data were excluded from the study.
Replication	We performed 18 independent biological replicates, by applying multiplexed single-cell reduced representation bisulfite sequencing (MscRRBS) to 6 different B cells healthy donors and 12 CLL patients. This translates into a total of 2,652 cells profiled by single-cell methylome sequencing. All attempts at replication were successful.
Randomization	Randomization is not applicable as no experimental groups were used in our study.
Blinding	Blinding is not applicable as no experimental groups were used in our study.

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

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

Antibodies

Antibodies used	The antibodies used for index sorting of normal B cells were: FITC Mouse Anti-Human IgD (Clone IA6-2, BD Pharmingen), APC Mouse Anti-Human IgG (#562025, BD Biosciences), APC/Cy7 anti-human IgM Antibody (Clone MHM-88, BioLegend) and PE/Cy7 anti-human CD27 Antibody (clone O323, Bio Legend). Antibody used for ChIP is anti-H3K27ac (2 mg for 25 mg of chromatin; ab4729 Abcam, Cambridge, United Kingdom).
Validation	Expression of Immunoglobulin Heavy Chain (IGH) genes was assessed by scRNAseq in index-sorted B cell subpopulations validating our index-sorting strategy (CD27-IgM+IgD+++IgG- [NBC], CD27-IgM+IgD+IgG- [loMBC], CD27+IgM+IgD++IgG- [intMBC], and CD27+IgG+ [hiMBC]). In addition, all antibodies used were validated for their use in FACS or ChIP-seq experiments with human samples, as shown on the website provided by the respective companies.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Relevant information on human research participants is provided in Figure 1b, Extended Data Figure 1b and Supplementary Table 12.
Recruitment	The diagnosis of CLL according to World Health Organization (WHO) criteria was confirmed in all cases by flow cytometry, or by

Recruitment	lymph node or bone marrow biopsy. IRB-approved protocols for genomic sequencing of patients' samples was obtained prior to the initiation of sequencing studies. Blood samples were collected in EDTA blood collection tubes from patients and healthy adult volunteers enrolled on clinical research protocols at the Dana-Faber/Harvard Cancer Center (DF/HCC) and NewYork-Presbyterian/Weill Cornell Medical Center (NYP/WCMC), approved by the DF/HCC and NYP/WCMC Institutional Review Boards.
Ethics oversight	The study was approved by the local ethics committee and by the Institutional Review Board (IRB) and conducted in accordance to the Declaration of Helsinki protocol. We note that the IRB does not permit collection of demographic information of healthy donors.

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

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119103
Files in database submission	GSM3358078 cll_175_h3k27ac_bam GSM3358082 cll_189_h3k27ac_bam GSM3358099 cll_248_h3k27ac_bam GSM3358103 cll_253_h3k27ac_bam
Genome browser session (e.g. UCSC)	No longer applicable

Methodology

Replicates	Two IGHV mutated and two IGHV unmutated CLL patient samples.
Sequencing depth	125bp paired-end mode. An average of 75 million paired reads was generated per sample
Antibodies	Antibody used for ChIP is anti-H3K27ac (2 mg for 25 mg of chromatin; ab4729 Abcam, Cambridge, United Kingdom).
Peak calling parameters	Peaks were identified with Macs2 (v2.0.10) with a q-value threshold of 0.01, according to the ENCODE Histone ChIP-seq Data Standards and Processing Pipeline (https://www.encodeproject.org/chip-seq/histone/).
Data quality	Deeptools plotFingerprint v2 was used to assess ChIP-seq signal enrichment over background signal. In addition, we observed a large overlap (72%) between FANTOM5 human robust enhancers (defined by H3K27ac signal) and our CLL H3K27ac ChIP-seq peaks, confirming the reproducibility of our ChIP-seq data.
Software	ChIP-seq data were processed according to the ENCODE Histone ChIP-seq Data Standards and Processing Pipeline (https://www.encodeproject.org/chip-seq/histone/). Raw reads were mapped to the human genome hg19 assembly using Burrows-Wheeler Aligner (BWA v0.7.17). Duplicate reads were removed using Picard (https://broadinstitute.github.io/picard/). Peaks were identified with Macs2 (v2.0.10) with a q-value threshold of 0.01. Peaks overlapping with Satellite repeat regions and Encode Blacklist were discarded.