nature research

Corresponding author(s): Jussi Taipale

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 for Data collection
Data analysis	The softwares used in this manuscript include MOODS (v1.9.1), MACS (v2.1.1), Trim Galore (v0.4.3), samtools (v1.3.1), bwa (v0.7.16a), BEDtools (v2.26.0), Autoseed (Jolma, A. et al., Nature 527, 384-388 (2015); Nitta, K. R. et al, Elife 4, e04837 (2015)), PEAR (v0.9.8), UCSF Chimera (1.11.2), R (v3.4.0) The R packages used include circlize (v0.4.3), ggplot2 (v2.2.1), FactoMineR (v1.39), NMF (v0.20.6) Custom codes are available on request, this is stated in "Code availability" of the Methods section

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

All next generation sequencing data are deposited to European Nucleotide Archive (ENA) under Accession PRJEB22684 and mentioned in Data Availability Statement.
Most of our analyses are based on the E-MI diagonal values, the raw data is provided in Suppl. Table S3.

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

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	The sequencing depth of NCAP-SELEX and HT-SELEX libraries was set to ensure that at least hundreds of thousands unique reads are available for each TF. Under this sample size, if a TF is binding nucleosomal DNA without restrictions, any non-random pattern of TF binding that has a biologically meaningful effect size (as observed in our study) can only occur with an extremely small p-value.
Data exclusions	The failed SELEX experiments were excluded according to the QC criteria. The criteria define successful TFs as having detectably stronger E-MI between neighboring 3-mer pairs than that between 3-mer pairs far away from each other, and showing enriched motifs that are not contaminations from unrelated TFs (see "Quality control of the SELEX experiments" section in Methods). The exclusion criteria is established before we perform conclusion-related analyses.
Replication	We performed multiple cycles (4–5) of SELEX for each TF. Each cycle is essentially a replicate of the same experiment. In addition, the whole SELEX procedure was also repeated for all TFs. For all the reported signals, their enrichment is observed across multiple SELEX cycles, and are reproducible between two or more independent batches of SELEX. Two replicates are available for the in vivo validation by MNase-ChIP.
Randomization	Samples were analyzed directly and individually, and not randomized to experimental groups
Blinding	Most analyses were performed using computational algorithms. Investigators were not blinded.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Antibodies

Antibodies used	The following antibody was used in this experiment: V5 Tag Monoclonal Antibody from Thermo Fisher Scientific, catalog # R960-25, RRID AB_2556564. The antibody was used at 1:200 dilution.
Validation	Validated in previous publications for Human and Mouse, and by the following techniques: Immunocytochemistry (ICC), Immunoprecipitation (IP), Western Blot (WB), Immunohistochemistry (IHC), ELISA, Flow Cytometry, ChIP assay (ChIP), Immunohistochemistry (Paraffin) (IHC (P)).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	LoVo (ATCC CCL-229), HEK293 (ATCC CRL-1573)
Authentication	The LoVo and HEK293 cells were directly obtained from the trusted vendor (ATCC) and not from other laboratories and used within short time.
Mycoplasma contamination	Tested to be free of mycoplasma infection by Hoechst staining
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Method-specific reporting

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging

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.ebi.ac.uk/ena/data/view/PRJEB22684
Files in database submission	Available in the project description of the linked database
Genome browser session (e.g. UCSC)	Available on request

Methodology

Replicates	Two technical replicates are available for each MNase-ChIP experiment. Peaks in replicates enrich similar motifs.
Sequencing depth	Total reads: 51285876 (RFX5) and 53017344 (HOXB13). Uniquely mapped reads: 9184303 (RFX5) and 9715845 (HOXB13). The sequencing run is paired-end. The sequencing length is 91 bp.
Antibodies	V5 Tag antibody (R96025, ThermoFisher). All information available on vendor website
Peak calling parameters	Peak called using MACS with -f BAMPE -g hs -B -q 0.1 --nomodel -m 2 50 for MNase-ChIP__cell-HEK293__TF-RFX5__Replicate-1.bam and MNase-ChIP__cell-HEK293__TF-HOXB13__Replicate-1.bam
Data quality	Correct TF motifs are discovered from the ChIP peaks. 6060 (HOXB13) and 1450 (RFX5) peaks are at FDR 5% and above 5-fold enrichment.
Software	bwa {Li H. and Durbin R., Bioinformatics 25, 1754-1760 (2009)} MACS {Zhang et al., Genome Biol. 9, pp. R137 (2008)}