nature research

Corresponding author(s): Merlin Crossley

☐ Initial submission ☒ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Since CRISPR/Cas-9 genome editing was such an important aspect of the experimental cell models established for this research, experiments were performed with CRISPR/Cas-9 genome edited clonal cell populations being derived from a parental clonal cell line. Two to three independent clonal populations (independent biological replicates; n=2-3) were derived for each HPFH mutation from the WT parental clone. In most cases, experiments were performed on three clonal populations (-114C>A, 13bp deletion and -195C>G), except in cases where it was difficult to introduce the HPFH mutation (-117G>A) and only 2 modified clones could be established. These samples numbers are consistent with the expectations of the field.

2. Data exclusions

Describe any data exclusions.

No data exclusions.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Experimental results were able to be reproduced in different clonal populations (n=2-3).

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Randomization was not applicable to this research as there was no group allocations since we were working with clonally derived cell populations.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Investigators were not blinded to sample identity as all data produced was from objective quantitative methods so subjective bias was not relevant.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

For ChIP-seq analysis:
Trimmomatic v0.3.2 was used for quality filtering and adapter trimming
Bowtie2 v2.2.9 was used for alignment.
Picard MarkDuplicates (http://broadinstitute.github.io/picard/) v1.109 was used to remove duplicate alignments.
The spp package v1.14 (https://github.com/kundajelab/phantompeakqualtools) was used for peak calling.
IDR2 (https://github.com/nboley/idr) was used for IDR analysis to obtain a set of consistent peaks.
MEME-ChIP v4.9.1 was used for motif analysis.
BEDtools (v2.26.0) genomecov for generating read coverage tracks.
HOMER v4.8 was used for peak annotation.
Peak centre to TSS plots were generated using GraphPad Prism version 7.0c
IGV Version 2.3.97 was used to visualise ChIP-seq read coverage tracks
For gene ontology analysis, two web-based software programs were used: DAVID Bioinformatics Resources 6.750,51 and GOrilla52,53 with the default human genome as background. GO terms were further processed by a web based software: REVIGO, to remove any redundant GO terms54.
ImageJ v1.51s was used to perform densitometry of EMSAs.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All unique materials are freely available.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

anti-BCL11A antibody (NB600_261, Novus Biologicals, lot: A2)
V5-tag (R960CUS, Life Technologies, lot: 1875853)
Beta-actin (A1978, Sigma)
ZBTB7A (Santa Cruz Biotechnology Inc, sc-33683x and eBioscience, 14-3309-82, lot: 4319430)
anti-goat IgG (sc-2028, lot: 12512)
anti-rabbit IgG (Cell Signaling Technologies #2729, lot: 8)
anti-FLAG M2 monoclonal antibody (Sigma F3165, lot: SLBT6752)

For ChIP 10-15 ug of anti-BCL11A antibody, ZBTB7A antibody or V5 antibody and respective IgG controls were used.

For EMSA supershift, 1 uL of anti-FLAG M2 monoclonal antibody or anti-BCL11A antibody was added to the reaction mix.

For Western blotting, V5 and Beta-actin antibodies were used at 1:15,000 dilution in TBST buffer.

Antibodies were validated for use in the system under study based on positive and negative controls. Positive and negative control loci were used to confirm the antibodies were working for ChIP assays and transiently over-expressed proteins from COS-7 cells confirmed that the antibodies recognise the appropriate proteins in Western Blots and EMSA super-shifts.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

COS-7 (Gift from Stu Orkin, Harvard)
K562 (Gift from Stu Orkin, Harvard)
HUDEP-2 (Ryo Kurita and Yukio Nakamura, Cell Engineering Division, RIKEN BioResource Center, Tsukuba, Ibaraki, Japan)

b. Describe the method of cell line authentication used.

None of the cell lines have been authenticated.

c. Report whether the cell lines were tested for mycoplasma contamination.

Cell lines were not tested for mycoplasma contamination.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Study did not involve human participants.

nature research

Corresponding author(s): Merlin Crossley

☐ Initial submission ☒ Revised version ☐ Final submission

ChIP-seq Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Data deposition

1. For all ChIP-seq data:

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

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

2. Provide all necessary reviewer access links. *The entry may remain private before publication.*	ChIP-Seq data has been deposited into the publically available database Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession number GSE103445
3. Provide a list of all files available in the database submission.	Track (bedGraph), peak and raw files are available for all ChIP-seq experiments. Modified genome used for HUDEP-2(delta-G-gamma) cells is also available.
4. If available, provide a link to an anonymized genome browser session (e.g. UCSC).	N/A

Methodological details

5. Describe the experimental replicates.	ZBTB7A ChIP-seq in K562 cells: 4 replicates (4 ChIP samples with respective input samples). ZBTB7A ChIP-seq in HUDEP-2 (delta-G-gamma) cells WT: 1 replicate with input ZBTB7A ChIP-seq in HUDEP-2 (delta-G-gamma) cells -195 C>G: 1 replicate with input BCL11A-ER-V5 ChIP-seq in HUDEP-2 cells: 2 replicates (2 ChIP samples with respective inputs)
6. Describe the sequencing depth for each experiment.	ZBTB7A ChIP-seq in K562 cells: 50bp single-end sequencing Replicate 1 - 42,964,189 reads, 95.27% mapped, 71.67% uniquely mapped Input 1 - 44,820,760 reads, 99.75% mapped, 73.13% uniquely mapped Replicate 2 - 52,954,533 reads, 96.53% mapped, 72.70% uniquely mapped Input 2 - 56,404,182 reads, 99.7% mapped, 73.40% uniquely mapped Replicate 3 - 51,065,734 reads, 97.73% mapped, 72.12% uniquely mapped Input 3 - 10,710,308 reads, 99.43% mapped, 57.47% uniquely mapped Replicate 4 - 55,346,970 reads, 98.49% mapped, 72.76% uniquely mapped Input 4 - 36,122,223 reads, 99.43% mapped, 64.86% uniquely mapped ZBTB7A ChIP-seq in HUDEP-2 (delta-G-gamma) cells WT: 76bp single-end sequencing Replicate 1 - 51,070,653 reads, 97.72% mapped, 75.04% uniquely mapped Input 1 - 56,612,731 reads, 99.09% mapped, 62.45% uniquely mapped ZBTB7A ChIP-seq in HUDEP-2 (delta-G-gamma) cells -195 C>G: 76bp single-end sequencing Replicate 1 - 53,971,108 reads, 98.03% mapped, 75.87% uniquely mapped Input 1 - 55,133,338 reads, 96.73% mapped, 51.68% uniquely mapped BCL11A-ER-V5 ChIP-seq in HUDEP-2 cells:

76bp paired-end sequencing
Replicate 1 - 55,568,296 reads, 99.03% mapped, 84.21% uniquely mapped
Input 1 - 63,891,446 reads, 99.42% mapped, 83.27% uniquely mapped
Replicate 2 - 66,009,714 reads, 99.08% mapped, 82.21% uniquely mapped
Input 2 - 71,222,398, 99.37% mapped, 82.12% uniquely mapped

7. Describe the antibodies used for the ChIP-seq experiments.

BCL11A-ER-V5 ChIP-seq in HUDEP-2 cells: antibody against V5-tag (R960CUS, Life Technologies).

ZBTB7A ChIP-seq in K562 cells: ZBTB7A antibody (Santa Cruz Biotechnology Inc, sc-33683).

ZBTB7A ChIP-seq in HUDEP-2(delta-G-gamma) cells: ZBTB7A antibody (eBioscience, 14-3309-82).

8. Describe the peak calling parameters.

After quality filtering and adapter trimming using Trimmomatic v0.3.2 (default parameters). Reads were aligned to human genome GRCh38 (hg38) using Bowtie v2.2.9 set to --very-sensitive. For the ZBTB7A ChIP-seq in HUDEP-2(delta-G-gamma) cells, reads were aligned to a modified version of GRCh38 which had a 4928bp deletion and a point mutation on chr11 between HBG1 and HBE. All downstream analysis for this dataset was done using the modified genome and an associated gene annotation file. These files will be provided on the NCBI GEO Submission.

Peak calling for all datasets was done using the spp package (https://github.com/kundajelab/phantompeakqualtools) on unique alignments (mapping quality (MAPQ) > 10) after removing duplicates with Picard MarkDuplicates (http://broadinstitute.github.io/picard/). The Input alignments were used as controls. Peak calling for the ZBTB7A ChIP-seq in K562 cells and the BCL11A-ER-V5 ChIP-seq was done by using the parameter -npeak=300000 to get a large number of relaxed peaks to then use for IDR analysis. For the ZBTB7A ChIP-seq in HUDEP-2(delta-G-gamma) cells, an FDR of 0.01 was used for peak calling. All other parameters for the spp package were default.

9. Describe the methods used to ensure data quality.

For the ZBTB7A ChIP-seq in K562 cells and the BCL11A-ER-V5 ChIP-seq, IDR analysis was performed (IDR=0.05) on initial peak lists using IDR2 (https://github.com/nboley/idr). This provided a conservative and optimal set of peak lists, the optimal peak list was used for further analysis. In the case of the ZBTB7A K562 ChIP-seq which had 4 replicates, IDR analysis was performed pairwise on all replicates and only the optimal peaks that were found across all pairwise replicate comparisons were used for further analysis.

10. Describe the software used to collect and analyze the ChIP-seq data.

Motif analysis was performed using MEME-ChIP on 200bp regions (+/-100bp) around the centre of the top 1000 peaks.
Peak lists were annotated using HOMER v4.8. For genomic localisation of peaks, promoters were considered as -1000bp to +100bp of the transcription start site (TSS). Peak centre to TSS plots were made using the annotated peak list with 20bp bins and a moving average curve (n=4) was fit to the data.

To further understand the gene ontology associated with BCL11A binding, we performed gene ontology term enrichment analysis with top 3000 BCL11A ChIP-Seq peaks and BCL11A ChIP-Seq peaks containing TGNCCA motif respectively. Two web-based software programs were used: DAVID Bioinformatics Resources 6.750,51 and GOrilla52,53 with the default human genome as background. GO terms were further processed by a web based software: REVIGO, to remove any redundant GO terms54.