Methods

Chromosome Conformation Capture (3C)

2-5 million cells were crosslinked with 2% formaldehyde for 15 minutes at room temperature. The cells were resuspended in 1 mL lysis buffer (10 mM Tris-HCl at pH 8.0, 10 mM NaCl, 0.2% NP-40) and incubated for 15 min. Cells were then resuspended in 0.5 mL 1.2x NEB CutSmart buffer, and shaken for 20 minutes at 65°C. Triton X-100 was added to 2% and nuclei were shaken for another 1 hour at 37°C. 400 U of EcoRI-HF (NEB, R3101) was added to the nuclei suspension and the reaction was carried out overnight at 37°C. The next day, 200 U EcoRI-HF was added and incubated for another 3 hrs. 88 µL 10% SDS was added to the nuclei, followed by 65°C incubation for 20 minutes to inactivate the restriction enzyme. To ligate the digested chromatin, the following components were added to the nuclei and brought up to 1.5 mL with distilled water: 0.15 mL 10x ligation buffer (NEB), 75 µL 20% Triton X-100, and 50 U T4 DNA ligase (ThermoFisher, EL0011). The reaction was carried out at 16°C for 4 hours and transferred to room temperature for another 30 minutes. DNA was purified by phenol extraction and isopropanol precipitation. Quantitative PCR was performed with iQ SYBR Green mix (Biorad). Ligation efficiencies were normalized to an independent ERCC3 locus. The Ct values for positive controls and interactions in a typical experiment ranged from 27-30. Primers are listed in Supplementary Table 2.

Construction of split Target-AID-NG

Introduction of the original Target-AID-NG produced protein at very low level (Extended Data Figure 4b) which did not lead to detectable base editing in HUDEP-2 cells. We reasoned that a higher level of base editor protein was needed for successful base editing. To this end, we engineered a split Target-AID-NG. The system had three components that are separately expressed (Extended Data Fig. 4a). The Cas9NG cDNA was amplified from pSI-Target-AID-NG[46] and inserted into pLVX lentiviral vector, followed by an SV40 nuclear localization signal, the N terminal part of Cfa intein, a 3xFlag tag and the C terminal part of mNG2[11]. The second component expresses mNG2[1-10], which is the N terminal part of split mNG2. When co-expressed with mNG2[11], they form a complete mNeonGreen molecule which facilitates cell sorting[78]. The other part of Target-AID-NG, composed of SH3-3xFlag-AID-SV40NLS-UGI, was inserted to pLenti vector followed by P2A-blasticidin.

The C terminal part of Cfa intein was inserted at the N terminus of SH3. All the three vectors were packaged into lentivirus.

To establish a cell line expressing full-length Target-AID-NG, we first expressed Cas9NG-NLS-CfaN-3xFlag-mNG2[11] and mNG2[1-10], and sorted cells using GFP gate. CfaC-SH3-3xFlag-AID-NLS-UGI-P2A-blast was then expressed and cells were selected with blasticidin. The complementary CfaN and CfaC will form a catalytically active intein that cleaves itself while ligates the two polypeptides, leaving three residues (CFN) in between. The full-length product, named split intein Target-AID-NG, contains a 3xFlag tag that facilitates its detection through immunoblot.

Chromatin Immunoprecipitation sequencing and qPCR

Five million differentiated HUDEP cells were collected and fixed with1% formaldehyde for 5 minutes at room temperature, and quenched with 125 mM glycine. Cells were washed with ice-cold PBS twice and resuspended in 0.13 mL lysis buffer (50 mM Tris-HCl 8.0, 10 mM EDTA, 0.5% SDS) and sonicated in a microtube (Covaris, 520045) with Covaris E220 ultrasonicator (Covaris) to an average of 300-500 bp. Sonicated chromatin was diluted with 1 mL ChIP dilution buffer (20 mM Tris-HCl 8.0, 2 mM EDTA, 1% Triton X-100, 300 mM NaCl, protease inhibitor), followed by adding 20 µL Dynabeads protein G (Thermo Fisher Scientific) and 3 µg antibody (NFYA, ab139402, Abcam). The mixture was incubated on a rotator overnight and washed with the following buffers: twice with RIPA150 (20 mM Tris-HCl 8.0, 1 mM EDTA, 1% Triton X-100,150 mM NaCl, 0.1% Sodium Deoxycholate, 0.1% SDS), twice with RIPA500 (20 mM Tris-HCl 8.0, 1 mM EDTA, 1% Triton X-100, 500 mM NaCl, 0.1% Sodium Deoxycholate, 0.1% SDS), twice with LiCl buffer (10 mM Tris-HCl 8.0, 1 mM EDTA, 0.5% NP-40, 250 mM LiCl, 0.5% Sodium Deoxycholate), twice with TE buffer (10 mM Tris-HCl 8.0, 1 mM EDTA). The washed chromatin was eluted using 300 µL Elution buffer (50 mM Tris-HCl 8.0, 10 mM EDTA, 1% SDS, 150 mM NaCl, 0.1 mg/mL Proteinase K) and decrosslinked at 65°C overnight. DNA was extracted with phenol-chloroform, and precipitated by adding 2.5x volume of absolute ethanol followed by centrifuging at 14,000 rpm for 15 minutes. The pellet was washed once with 75% ethanol, then dried and dissolved with 50 µL TE buffer. To construct ChIP-seq library, we used NEBNext Ultra II DNA Library Prep Kit (NEB) according to manufacturer's protocol. Quality check of libraries were carried out with Qubit and bioanalyzer. The libraries were sequenced on the NextSeq 500 platform with NextSeq 500/550 High Output Kit v2 (75 cycles). Paired-end sequencing was performed (2x42 bp, 6 bp index). qPCR validation was performed using 1 µL ChIP

product as template. ChIP-qPCR primers for γ-globin and β-globin promoters are listed in Supplementary Table 2.

ChIP-seq data analysis

ChIP-seq data were quality checked with FASTQC and processed with CUT&RUNTools, which is compatible with all paired-end sequencing technologies including ChIP-seq. CUT&RUNTools filters and remove DNA fragments that are larger than 120 bp as a default. This setting was changed to removing >500 bp DNA. Other parameters were kept unchanged. Peaks were called using MACS2 with the following parameters: -g hs -f BAMPE -q 0.01 -B –SPMR. The resulting bigwig files were visualized with IGV. To compare between samples, deeptools were used to generate comparison matrix and plot heatmaps.

ATAC-seq and data analysis

Omni-ATAC[79] was performed with 50,000 fresh cells or frozen pellets. Cells were first permeabilized with 50 µL RSB buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$) with 0.1% NP40, 0.1% Tween-20 and 0.01% Digitonin for 3 minutes. 1 mL of cold RSB containing 0.1% Tween-20 was added before cells were pelleted again. The pellets were then resuspended with 50 µL transposition mixture (25 µL 2xTD buffer, 2.5 µL transposase from ATAC-seq kit (Illumina), 16.5 µL PBS, 0.5 µL 1% digitonin, and 0.5 µL 10% Tween-20), and reaction was carried out at 37 °C for 60 minutes. Genomic DNA was extracted with Qiagen MinElute PCR purification kit according to manufacturer's manual. Total DNA was used as template for amplification, mixed with 15 µL of 2x Ultra II Q5 mix, 1 µL of 10 µM i7 primer (AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGA-TGTG) and 1 µL of 10 µM i5 primer (CAAGCAGAAGACGGCATACGAGATNNNNNNNN-GTCTCGTGGGCTCGGAGATGT) where N represent barcodes. PCR amplification was performed as follow: 65°C 5 min, 98°C 30 s, 98°C 10 s, 63°C 30 s, repeat step 3-4 for 9 times, and 63°C 5min. PCR products were purified with AMPure beads and quantified using Qubit and Tapestation before pooling and sequencing. The libraries were sequenced in the NextSeq 500 platform with NextSeq 500/550 High Output Kit v2 (75 cycles). Paired-end sequencing was performed (2x42 bp, 6 bp index). ATAC-seq data was processed using CUT&RUNTools using same setting with CUT&RUN data. The output bigwig files were visualized using IGV.

PRO-seq

PRO-seq was performed as described[80]. Ten million $CD34^+$ cells undergoing erythroid differentiation were harvested on ice, and washed with cold PBS. Cells were resuspended in 2 mL cold Buffer W (10 mM Tris-Cl, pH 8.0, 10 mM KCl, 250 mM sucrose, 5 mM $MgCl_2$, 0.5 mM DTT, 10% glycerol, 0.5% BSA) using wide-bore P1000 tip and transferred to a 2 mL low binding tube, and strained through a 35 micron nylon mesh cell strainer. Immediately add 18 mL Chilled Buffer P (Buffer W supplemented with 0.05% Igepal CA-630) to cell suspension and mix gently for 2 min. Cells were centrifuged at 400 x *g* for 8 min and resuspended in 1 mL Buffer F (50 mM Tris-HCl pH 8.3, 40% glycerol, 5 mM $MgCl_2$, 0.5 mM DTT, 0.5% BSA), and centrifuged again at 400 x *g* for 8 min. The pellets were resuspended in 400 µL Buffer F, counted using a hemocytometer to ensure that >95% cells were permeabilized. The cell density was adjusted to 1 x 10^6/100 µL, and the suspension was snap-frozen and stored at -80°C.

An aliquot of frozen cells was thawed on ice and counted. For each sample, take 1 million cells and spike in 50,000 permeabilized Drosophila S2 cells, and add the same volume of 2x Nuclear Run-On buffer (10 mM Tris-HCl pH 8.0, 300 mM KCl, 1% Sarkosyl, 5 mM $MgCl_2$, 1 mM DTT, 200 µM biotin-11-A/C/G/UTP (Perkin-Elmer), 0.8 u/mL SUPERaseIN inhibitor [Ambion]) and then incubate at 30°C for 5 min. Total RNA was isolated and fragmented by base hydrolysis with 0.25 N NaOH on ince for 10 min, and neutralized with 1x volume of Tris-HCl pH 6.8. Biotinylated RNA was bound to 30 µL Streptavidin M-280 beads (Thermo Fisher Scientific) in Binding Buffer (300 mM NaCl, 10 mM Tris-HCl pH 7.4, 0.1% Triton X-100). Beads were washed twice with High salt buffer (2 M NaCl, 50 mM Tris- HCl pH 7.4, 0.5% Triton X-100), twice in Binding Buffer, and twice in Low salt buffer (5 mM Tris-HCl pH 7.4, 0.1% Triton X-100). Bound RNA was extracted using Trizol (Invitrogen), precipitated using ethanol and dissolved in H_2O. Nascent RNA was ligated with 10 pmol of 3' RNA adaptor (5'p-rNrNrNrNrNrGrArUrCrGrUrC-rGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC-/3'InvdT/) and T4 RNA ligase I (NEB) at 20°C for 2hrs. Ligated RNA was enriched with another round of streptavidin beads purification with one additional round of washing with 1x ThermoPol Buffer (NEB). RNA products were treated with RNA 5' Pyrophosphohydrolase (RppH, NEB) at 37°C for 30 min followed by one wash of High, Low and T4 PNK Buffer. 5' ends were repaired with Polynucleotide Kinase (PNK, NEB) at 37°C for 30 min. RNA was ligated to 5' RNA adaptor 5'-rCrCrUrUrGrGrCrArCrCrCrGrArGrArArUrUrCrCrA-3' with T4 RNA ligate I (NEB) for 2 hrs at 20 °C, and enriched with streptavidin beads with additional wash using 1x SuperScript IV buffer (Thermo Fisher Scientific) and reverse transcribed using 25 pmol RT

primer AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA. 12.5 pmol RPI index primers and RT pimer were added to the RT product for library amplification. Libraries were purified and size selected, and sequenced using Nextseq500 platform. Paired-end sequencing, read length 42 bp x 2, 6 bp index was performed.

PRO-seq data processing

FASTQ read pairs were trimmed to 40 bp per mate, and read pairs with a minimum average base quality score of 20 were retained. Read pairs were then further trimmed using cutadapt 1.14 to remove adapter sequences and low-quality 3' bases (--match-read-wildcards -m 20 -q 10).

R1 reads were then aligned to the drosophila spike genome (index using Bowtie 1.2.2 (-v 2 --best --un), with those reads failing to map serving as input to the primary genome (hg19) alignment step(s) (using Bowtie options -v 2 --best). Reads mapping to the reference genome were sorted via samtools 1.3.1 (-n), and subsequently converted to bedGraph format. Because PRO-seq reveals the position of the RNA 3' end, the "+" and "-" strands were swapped to generate bedGraphs representing 3' end positions. TSS-centric read count matrices were then calculated over a window of +/- 2kb with 25bp bins.

A table of statistics was compiled across all samples, summarizing mapping performance, TSS-proximal (TSS +/- 150) read count correlation coefficients, and condition-level spike differences. Sample-level metagene and PCA figures were also generated. For both TSS-proximal read count correlations and PCA analyses, TSS's with 0 binned reads across all samples were excluded.

Refgene annotations were obtained from assembly GCF_000001405.25 for GRCh37. The strand specific bed files from PRO-seq were mapped to the genes with the same transcription direction. The total read count was calculated for each gene, and was then normalized by the total number of reads in the sample and reported as reads per million. The normalized reads for each gene were compared across *BCL11A* KO samples and *AAVS1* controls, and the log2 fold change was calculated for both 32 hrs and 72 hrs samples. Genes with log2 fold change larger than one was reported.