Article

Mitochondrial base editor induces substantial nuclear off-target mutations

https://doi.org/10.1038/s41586-022-04836-5

Received: 10 May 2021

Accepted: 5 May 2022

Published online: 12 May 2022

Check for updates

Zhixin Lei[1,2,9], Haowei Meng[3,9], Lulu Liu[3,9], Huanan Zhao[4,5,9], Xichen Rao[3], Yongchang Yan[1,2], Hao Wu[1,2], Min Liu[1,6], Aibin He[1,6] & Chengqi Yi[1,3,7,8]✉

DddA-derived cytosine base editors (DdCBEs)—which are fusions of split DddA halves and transcription activator-like effector (TALE) array proteins from bacteria—enable targeted C•G-to-T•A conversions in mitochondrial DNA[1]. However, their genome-wide specificity is poorly understood. Here we show that the mitochondrial base editor induces extensive off-target editing in the nuclear genome. Genome-wide, unbiased analysis of its editome reveals hundreds of off-target sites that are TALE array sequence (TAS)-dependent or TAS-independent. TAS-dependent off-target sites in the nuclear DNA are often specified by only one of the two TALE repeats, challenging the principle that DdCBEs are guided by paired TALE proteins positioned in close proximity. TAS-independent off-target sites on nuclear DNA are frequently shared among DdCBEs with distinct TALE arrays. Notably, they co-localize strongly with binding sites for the transcription factor CTCF and are enriched in topologically associating domain boundaries. We engineered DdCBE to alleviate such off-target effects. Collectively, our results have implications for the use of DdCBEs in basic research and therapeutic applications, and suggest the need to thoroughly define and evaluate the off-target effects of base-editing tools.

Mutations in mitochondrial DNA (mtDNA) are known to be associated with most adult-onset mitochondrial diseases, which affect up to about 1 in 5,000 adults[2–5]. Although several severe syndromes related to mtDNA mutations have been reported, there are few effective treatments and no known cure[2,6]. Various gene therapy strategies have been developed to address this challenge[2]. For instance, mitochondrion-targetable nucleases such as zinc-finger nucleases (ZFNs) and TALE nucleases (TALENs) have been used to reduce the level of heteroplasmy in cells through direct degradation of mutated mtDNA molecules[7–9]. More recently, RNA-free DddA-derived cytosine base editors (DdCBEs) have been reported to precisely edit mtDNA without causing double-stranded breaks[1]. Thus, unlike the previous destruction-based strategies, this approach does not pose a risk of reducing the copy number of mtDNA to harmfully low levels, especially for cases of high mutation load.

The mitochondrial base editor is based on $DddA_{tox}$, a bacterial toxin that catalyses the conversion of deoxycytosine (dC) to deoxyuracil (dU) on double-stranded DNA[1] (dsDNA). To avoid potential toxicity, the deaminase has been split into two inactive halves, one containing the N terminus of $DddA_{tox}$ ($DddA_{tox}$-N) and the other containing the C terminus ($DddA_{tox}$-C). These halves reconstitute deamination activity when assembled by a pair of mitochondrial targeting signal (MTS)-linked TALE proteins, in a manner similar to the assembly of FokI monomers in zinc-finger nucleases and TALENs. Therefore, DdCBEs induce the intended dC-to-dU edits only when the two TALE repeats bind simultaneously to the on-target genomic sites.

Although DdCBE is a promising approach for treatment of mitochondrial diseases, unbiased and comprehensive analyses of its off-target effects are still lacking. Mok and colleagues reported varying degrees of off-target edits in mtDNA and no off-target effect in the nuclear DNA (nDNA), based on their analysis of the nuclear pseudogenes[1]; however, the genome-wide specificity of DdCBE remains unaddressed.

Assessing DdCBE specificity via dU

DdCBEs catalyse dC-to-dU conversions and finally result in dC-to-dT transitions. We recently developed an unbiased specificity evaluation method, Detect-seq[10], which is based on chemical labelling and enrichment of dU generated in vivo[11–13]. Using this method, we profiled the editome of cytosine base editors (CBEs)[14–18], and found unexpected off-target edits outside of protospacer and on the target strand[10]. Because DdCBEs rely on the same intermediate dU to achieve base editing in mtDNA, we aimed to apply Detect-seq to evaluate the genome-wide specificity of DdCBEs (Fig. 1a and Extended Data Fig. 1).

We transfected $DddA_{tox}$ fusions comprising the C-terminal half of $DddA_{tox}$ split at G1397 and bound to the right TALE assembled with the N-terminal half of $DddA_{tox}$ split at G1397 and bound to the left TALE[1] (Right–G1397-C + Left–G1397-N (hereafter abbreviated to L1397-N)) into HEK293T cells to target the mitochondrial genes *ND6*, *ND5* and *ND4*, forming *ND6*-L1397-N, *ND5.1*-L1397-N and *ND4*-L1397-N, respectively.

[1]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China. [2]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China. [3]State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, China. [4]School of Life Sciences, Tsinghua University, Beijing, China. [5]Peking University–Tsinghua University–National Institute of Biological Sciences Joint Graduate Program, School of Life Sciences, Tsinghua University, Beijing, China. [6]Beijing Key Laboratory of Cardiometabolic Molecular Medicine, Institute of Molecular Medicine, College of Future Technology, Peking University, Beijing, China. [7]Department of Chemical Biology and Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. [8]Peking University Genome Editing Research Center, Peking University, Beijing, China. [9]These authors contributed equally: Zhixin Lei, Haowei Meng, Lulu Liu, Huanan Zhao. ✉e-mail: chengqi.yi@pku.edu.cn

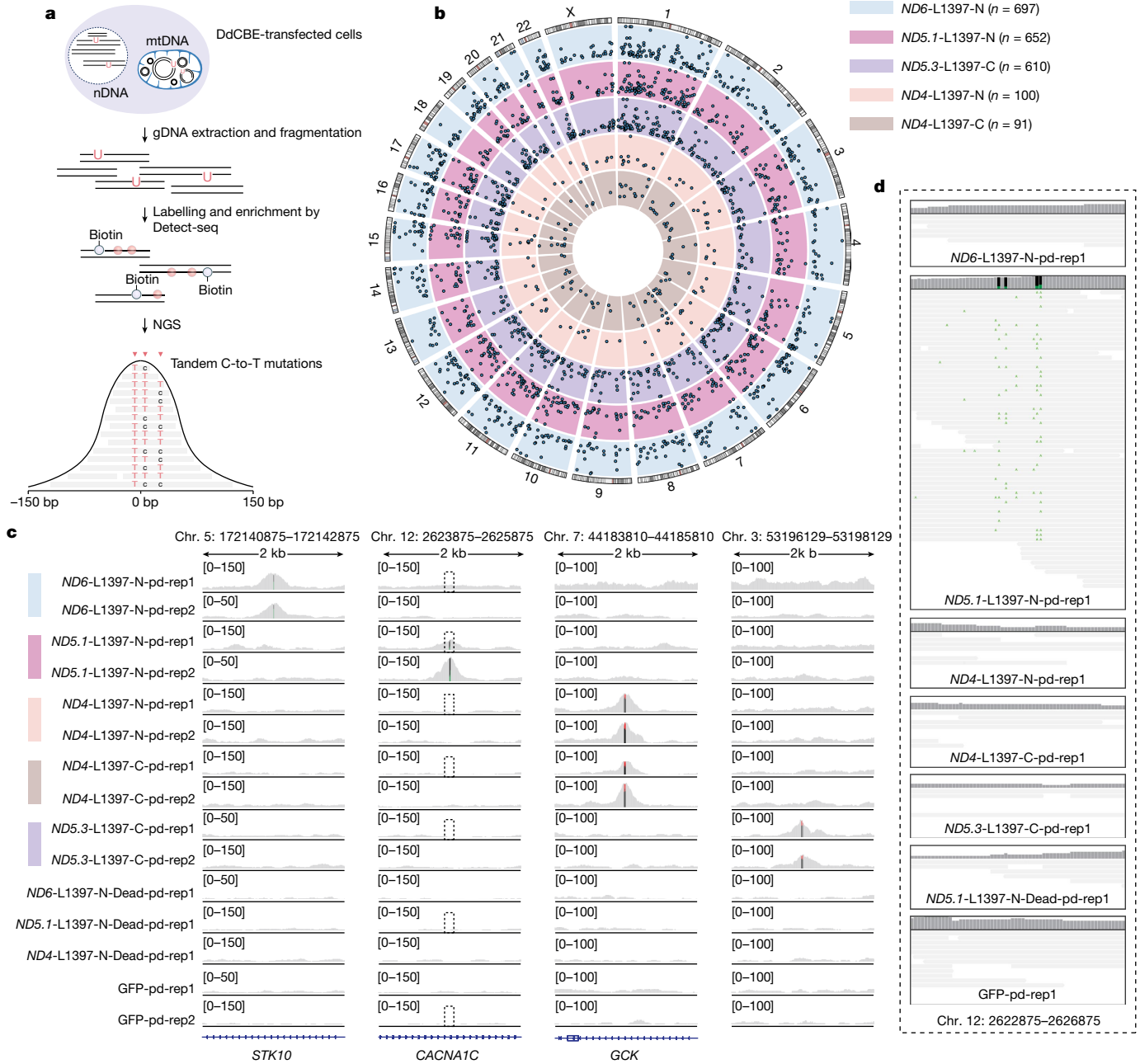


Fig. 1 | DdCBE induces abundant off-target edits in the nuclear genome. **a**, Overview of the use of Detect-seq to identify genome-wide off-target edits by DdCBE. DdCBE-treated cells were collected three days after transfection and genomic DNA (gDNA) was extracted. Deoxyuridine (dU) generated by DdCBE in both nuclear and mtDNA was recognized by uracil DNA glycosylase (UDG) and labelled with biotin and a mutagenic cytosine analogue. The labelled fragments were enriched through biotin pulldown for next generation sequencing (NGS); the cytosine analogues induce a tandem C-to-T mutation pattern to trace the editing events of DdCBE. **b**, Genome-wide circos plots representing the distribution and Detect-seq scores of identified nDNA off-target sites on each chromosome for the five G1397-split DdCBEs. **c**, Detect-seq results at four representative off-target sites identified for the G1397-split DdCBEs. **d**, Zoomed-in Integrative Genomics Viewer (IGV) views for the representative Detect-seq results (marked by the dashed boxes) in **c**. G•C-to-A•T mutations are indicated in green, and the green and black bars represent the ratio of G•C-to-A•T mutations at the sites. Rep, replicate; pd, pull-down.

Similarly, we transfected fusions of the C-terminal half of $DddA_{tox}$ split at G1397 and bound to the left TALE assembled with the N-terminal half of $DddA_{tox}$ split at G1397 and bound to the right TALE (hereafter abbreviated to L1397-C) into HEK293T cells to target the mitochondrial genes *ND5* and *ND4*, forming *ND5.3*-L1397-C and *ND4*-L1397-C, respectively. These DdCBEs achieved high on-target editing efficiencies (Supplementary Fig. 2). We then applied Detect-seq to profile the editome of DdCBE. As expected, we observed characteristic Detect-seq signals at the on-target sites (Supplementary Fig. 3). Previous work reported different levels of mtDNA off-target effects for the five DdCBEs[1]; consistent with these results, we observed the highest and lowest average level of mtDNA-wide off-target Detect-seq signals for *ND6* and *ND4* DdCBEs, respectively (Supplementary Fig. 4). Thus, we find that Detect-seq faithfully recapitulates the editome of DdCBEs in mtDNA.

Nuclear off-target editing by DdCBEs

We next performed unbiased analysis of Detect-seq results in the nuclear genome. We found 697, 652 and 100 off-target sites in nDNA for *ND6*-L1397-N, *ND5.1*-L1397-N and *ND4*-L1397-N, respectively, as well

as 610 and 91 off-target sites for *ND5.3*-L1397-C and *ND4*-L1397-C, respectively (Fig. 1b and Supplementary Table 1). Similarly large numbers of nuclear off-target edits were observed with different transfection protocols (Extended Data Fig. 2). Control cells expressing only GFP or DdCBEs with a catalytically inactive DddA resulted in background levels of Detect-seq signals at these sites (Fig. 1c,d and Supplementary Figs. 5 and 6). We then selected 65, 75 and 54 off-target sites for *ND6*-L1397-N, *ND5.1*-L1397-N and *ND4*-L1397-N, respectively, and verified them using targeted deep sequencing (Extended Data Fig. 3 and Supplementary Tables 2 and 3); these sites were selected to reflect sites with high, middle and low Detect-seq signals (Supplementary Tables 1 and 2). All 194 of the selected sites were validated as genuine off-target nDNA editing sites, with average editing ratios of approximately 3.18%, 2.31% and 0.46% for *ND6*-L1397-N, *ND5.1*-L1397-N and *ND4*-L1397-N, respectively; the most severe off-target mutations among the tested sites had editing ratios of 13.43%, 17.51% and 2.78%, respectively (Extended Data Fig. 3 and Supplementary Tables 2 and 3). In addition, under different transfection conditions, we deep sequenced 69 further off-target loci; we validated all of the 69 off-target sites, which suggested similar off-target effects under similar on-target efficacies (Supplementary Fig. 7 and Supplementary Tables 2 and 3).

To support the induction of off-target editing by DdCBE in the nuclear genome, we examined the distribution of *ND6*-L1397-N in different subcellular fractions by western blot and immunofluorescence. Whole-cell immunofluorescence showed that DdCBE was preferentially localized within mitochondria (Supplementary Fig. 8), consistent with previous results[1]; the strong signals in mitochondria may interfere with the analysis of potential nuclear-localized DdCBE. To examine whether a small proportion of DdCBE could be aberrantly localized to cell nuclei, we used a non-fixation immunofluorescence strategy[19] to isolated nuclei from HeLa cells with confirmed, predominant mitochondrial localization of DdCBE. This assay maintained the 3D structure of isolated nuclei, enabling us to calculate the fluorescence intensity inside each nucleus (two examples are shown in Supplementary Videos 1 and 2). Nuclei from cells transfected with DdCBE showed significantly higher fluorescence intensity than those transfected with vector controls, regardless of the transfection conditions (Extended Data Fig. 4). We also performed western blot and immunofluorescence of fixed HEK293T cells to support the presence of DdCBE in the nucleus. We used a nuclear–cytosol cell fractionation assay and showed that in addition to the presence of *ND6*-L1397-N in the cytoplasmic fraction, we also observed DdCBE in the chromatin fraction (Extended Data Fig. 5a and Supplementary Fig. 9). The TALE arrays had a strong influence on nuclear localization compared with DddA or uracil DNA glycosylase inhibitor (UGI) (Extended Data Fig. 5). These western blot and immunofluorescence results were recapitulated using different transfection reagents and protocols (Extended Data Figs. 4 and 5). Finally, we performed in situ chromatin immunoprecipitation with sequencing (ChIP–seq) experiments[20] to reveal potential binding sites of *ND6*-L1397-N in the nuclear genome (Supplementary Fig. 10a). The ChIP–seq signals were highly correlated with the 697 off-target editing sites (Supplementary Fig. 10b). One-hundred and thirty seven out of the 697 off-target editing sites overlapped directly with DdCBE-enriched peaks (out of a total of 20,983 DdCBE-enriched peaks; Supplementary Table 4), representing a significant enrichment (chi-squared test P-value $< 2.2 \times 10^{-16}$) over the genomic background (Supplementary Fig. 10c).

TAS-dependent nuclear off-target edits

To examine how these off-target sites were generated, we systematically performed Detect-seq for *ND6*-L1397-N, *ND5.1*-L1397-N and *ND4*-L1397-N constructs lacking either the left or right TALE array or lacking both TALE arrays (Fig. 2a). We found 84 and 33 sites were sensitive to the depletion of left and right TALE arrays of *ND6*-L1397-N, respectively; 32 and 91 sites were sensitive to the depletion of left and right TALE arrays of *ND5.1*-L1397-N, respectively; and 0 and 30 sites were sensitive to the depletion of left and right TALE arrays of *ND4*-L1397-N, respectively (Fig. 2b, c and Supplementary Table 1). These sites were identified only for their respective DdCBE, meaning that they are specific for that TALE array (Fig. 1c). By contrast, we did not find any off-target edits that require the presence of both TALE arrays (Fig. 2c). This observation conflicts with the design principle of DdCBE, in which the editing activity is dependent on the reassembly of $DddA_{tox}$ halves at a genomic locus specified simultaneously by both by the left and right TALE arrays. To further validate that the TAS dependence of these off-target sites is unilateral, we performed targeted deep sequencing of genomic DNA edited by various *ND6*-L1397-N deletion constructs (Fig. 2a). Indeed, the results confirmed that such off-target editing is dependent on only one of the two TALE repeats (Fig. 2d, Supplementary Fig. 11 and Supplementary Tables 2 and 3). We also ruled out the possibility that these off-target sites are induced by only one TALE-bound $DddA_{tox}$ N-terminal or $DddA_{tox}$ C-terminal half without forming an intact deaminase.

To further understand the unilateral TAS dependence, we searched for putative TALE array binding sites (pTBS) among these off-target sites. Using the three L1397-N DdCBEs as examples, we identified a single pTBS for each of the TAS-dependent off-target sites (Supplementary Fig. 12); these pTBSs are located adjacent to the Detect-seq signals with their 3′ ends usually located around 4–11 bp away from the edits (Extended Data Fig. 6), in agreement with the preferred editing distance of G1397-split DdCBEs[1]. Inspection of the aligned pTBSs reveal frequent G-to-A mismatches (Fig. 2e, Extended Data Fig. 6b and Supplementary Fig. 12), which could be explained by the high affinity of the repeat-variable diresidue (RVD) NN for both A and G[21,22]. In addition, the N-terminal domain (NTD) of the right TALE of *ND6*-L1397-N for the N_0 position was engineered to be permissive for A, T, C and G nucleotides[1,23]; we obtained consistent observations, with T being slightly preferred in the off-target sites (Fig. 2e and Supplementary Fig. 12a). Taking the above factors into consideration, the aligned pTBSs show high similarity to the on-target site, containing no more than three mismatches with the binding sequence of either the left or right TALE array (Supplementary Fig. 12). We did not find plausible paired pTBSs for the vast majority of the off-target sites on the basis of sequence similarity, TALE orientation and spacing region length. Therefore, our in silico pTBS analysis supports the experimentally determined unilateral TAS dependence by Detect-seq and targeted deep sequencing (Fig. 2c–e, Extended Data Fig. 6b and Supplementary Fig. 12). The TAS-dependent nuclear off-target sites are probably caused by the spontaneous assembly of split $DddA_{tox}$ halves at the genomic loci and guided by one TALE array. Thus—in contrast to the design principle—one TALE array is sufficient to specify the off-target sites.

TAS-independent off-target edits in nDNA

Further analysis of Detect-seq results for various deletion constructs of *ND6*-L1397-N, *ND5.1*-L1397-N and *ND4*-L1397-N also revealed 542, 454 and 53 nuclear off-target sites that are independent of their TAS, respectively (Figs. 2a and 3a and Supplementary Table 1). Detect-seq signals of these sites remain strong upon depletion of either one or both of the paired TALE arrays (Fig. 3a,b). In addition, we validated the TAS independence of the sites by targeted deep sequencing (Extended Data Fig. 7 and Supplementary Tables 2 and 3). Nevertheless, the off-target sites are dependent on the intact deaminase, as depletion of either $DddA_{tox}$ half leads to a complete loss of editing activity at these sites. Moreover, we searched for pTBS within genomic loci of the off-target sites and could not identify plausible pTBS pairs for any of the TAS-independent off-target sites; out of the 542 sites for *ND6*-L1397-N, we could not identify pTBS for either the left or the right TALE array for 520 or 524 sites, respectively. The 7–9% of remaining sites with a single pTBS may be explained by the comparatively short (10- and 11-nt) TALE arrays of *ND6*-L1397-N, whose recognition sequences occur frequently

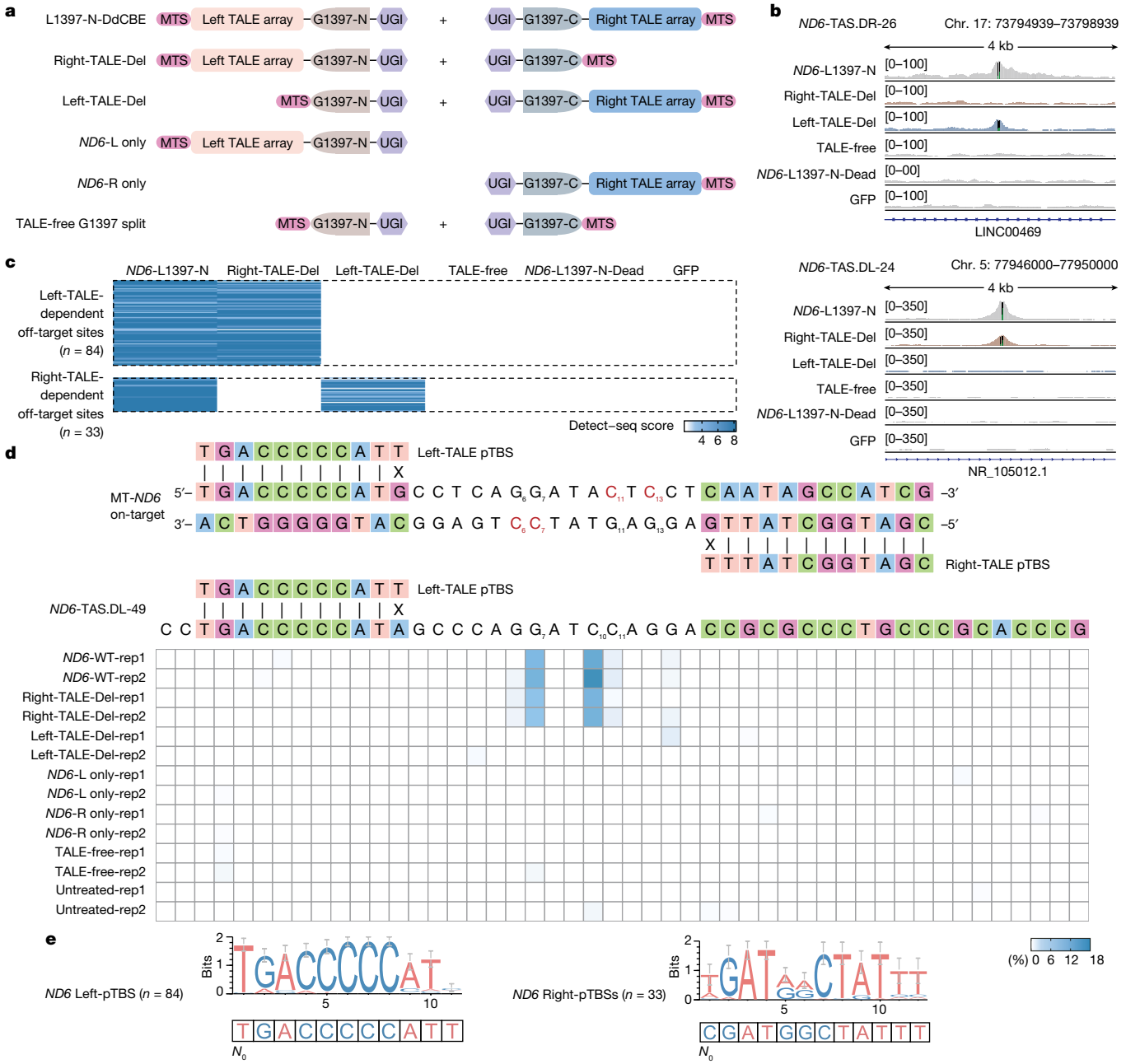


Fig. 2 | DdCBE induces one-sided, TAS-dependent nDNA off-target edits. **a**, DdCBE constructs used in the experiments. Related plasmids were constructed from full-length *ND6*-L1397-N, *ND5.1*-L1397-N and *ND4*-L1397-N. G1397-N and G1397-C are the N- and C-terminal halves of $DddA_{tox}$ split at G1397. **b**, Detect-seq results at representative right-TAS-dependent (top) or left-TAS-dependent (bottom) nDNA off-target sites for different *ND6*-L1397-N constructs. Dead indicates a catalytically inactive mutant of DddA. **c**, Detect-seq signals of all *ND6*-L1397-N TAS-dependent nDNA off-target sites for the constructs in **a**. Data are grouped into left-TAS- and right-TAS-dependent off-target sites. **d**, Editing ratio obtained from targeted deep sequencing at a representative left-TAS-dependent off-target site for different *ND6*-L1397-N constructs and untreated cells. The alignment of the on-target site is presented at the top, and the putative TALE array binding site (pTBS) is shown above the genome sequence. **e**, Sequence logos generated from the pTBS sequences from the left- and right-TAS-dependent off-target sites of *ND6*-L1397-N. Bits reflect the level of sequence conservation at a given position. The designed binding sequences are shown at the bottom.

throughout the genome (Supplementary Fig. 13). In total, we identified 744 TAS-independent off-target sites for the five DdCBEs.

An unexpected observation is that a majority (569 out of 744) of the TAS-independent nuclear off-target sites are shared by at least two of the five DdCBEs differing in TALE arrays and fusion orientation (Fig. 3c,d and Supplementary Figs. 14 and 15). The remaining 175 sites did not pass the threshold for significance for more than one DdCBE, but all demonstrated clear Detect-seq signals for at least one other DdCBE (Supplementary Figs. 16 and 17). To validate that TAS-independent off-target sites could be shared among different DdCBEs, we selected 12 shared sites and performed targeted deep sequencing for each sample transfected by the five DdCBEs. The results show that all of these sites were indeed edited in cells containing any DdCBE, with average editing ratios of about 6.68%, 4.85%, 0.44%, 3.94% and 0.42% for *ND6*-L1397-N, *ND5.1*-L1397-N, *ND4*-L1397-N, *ND5.3*-L1397-C and *ND4*-L1397-C, respectively (Fig. 3e,f and Supplementary Tables 2 and 3). We also interrogated the 12 sites for 2 additional DdCBEs with different TALE arrays or DddA split (*ND1*-L1397-N and the C-terminal half of $DddA_{tox}$ split at G1333 and

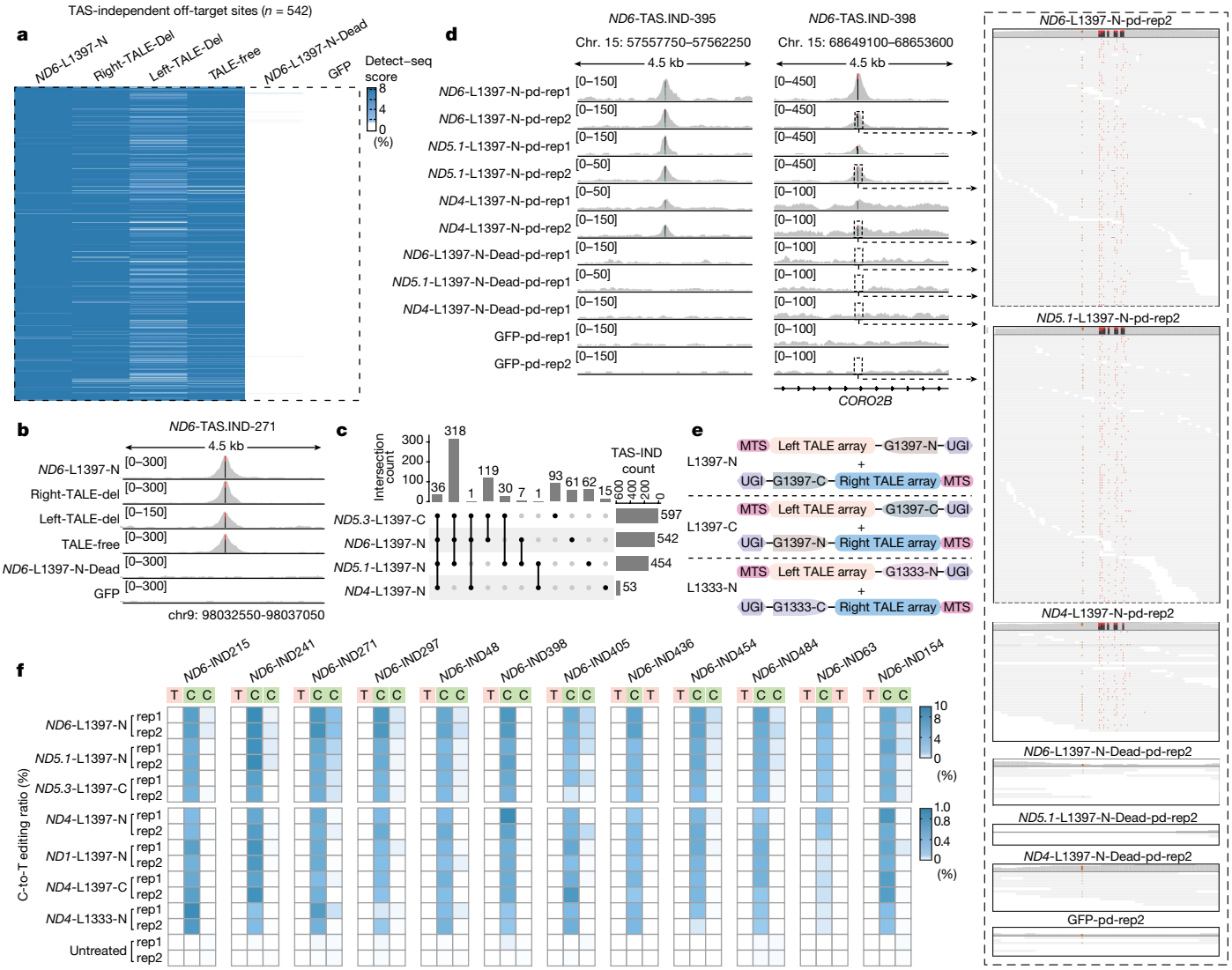


Fig. 3 | Prevalent, non-random TAS-independent off-target sites. **a**, Detect-seq signals for the *ND6*-L1397-N constructs in Fig. 2a at the TAS-independent nDNA off-target sites, which are insensitive to the deletion of TALE arrays. **b**, Detect-seq results at a representative TAS-independent off-target site for different *ND6*-L1397-N constructs. **c**, Upset plot showing the overlap of TAS-independent nDNA off-target sites for four G1397-split DdCBEs with distinct TALE arrays. **d**, Detect-seq results at two representative TAS-independent off-target sites identified for the three L1397-N DdCBEs. The bars indicate the ratio of G•C-to-A•T (green to black) or C•G-to-T•A (red to black) mutations at the sites. The zoomed-in IGV views for the Detect-seq data marked by the dashed boxes are shown on the right, in which the C•G-to-T•A mutations are indicated in red and A•T-to-G•C single nucleotide variants are indicated in brown. Reads over 200 are omitted owing to space limitations. The full image is shown in Supplementary Fig. 14. **e**, Additional DdCBE constructs used in the experiments. **f**, Editing ratio from targeted deep sequencing at 12 selected TAS-independent off-target sites for different constructs in **e**. No obvious pTBSs could be found at these sites; and the C nucleotides with the highest editing ratios are shown.

bound to the right TALE assembled with the N-terminal half of $DddA_{tox}$ split at G1333 and bound to the left TALE targeting *ND4* (*ND4*-L1333N)). These sites were also edited by these constructs, with average editing ratios of about 0.43% and 0.28%, respectively. Thus, unlike cytosine base editor (CBE)-induced Cas9-independent off-target sites[24–26], DdCBE-induced TAS-independent off-target sites are non-random and exhibit comparatively high editing ratios.

An encounter with CTCF

We next sought to investigate the nature of the TAS-independent off-target sites. As expected, we observed high DNase-seq signal at the TAS-independent off-target sites, suggesting a preference for open chromatin regions (Supplementary Fig. 18 and Methods). We also analysed histone mark signals at these sites and found relatively weak correlations with several active chromatin marks (Supplementary Fig. 19, Methods and Supplementary Discussion). Because many off-target sites are universally induced by different DdCBEs, we searched for potential common features by analysing the sequence context flanking the off-target regions. As expected, we observed an evident 5′-TC–3′ motif (where the underlined C indicates the modified cytosine) for the C nucleotides with the highest Detect-seq signals (Fig. 4a), consistent with the known sequence preference of the deaminase[1]. Further, we observed a strong GC-rich motif 8–9 bp downstream of the TC motif for 618 out of the 744 TAS-independent off-target sites (Fig. 4b, Extended Data Figs. 7 and 8 and Supplementary Fig. 20). This consensus sequence matches very well with the 12-bp core binding motif of the CTCF protein[27,28] (Fig. 4b, Extended Data Fig. 8 and Supplementary Fig. 20).

CTCF is a well-known factor with a role in organizing the 3D genome architecture, and forms loop domains in a process involving the cohesin complex[29,30]. To determine whether these off-target regions are indeed CTCF binding sites, we analysed the ChIP–seq data for CTCF, SMC3[31,32]

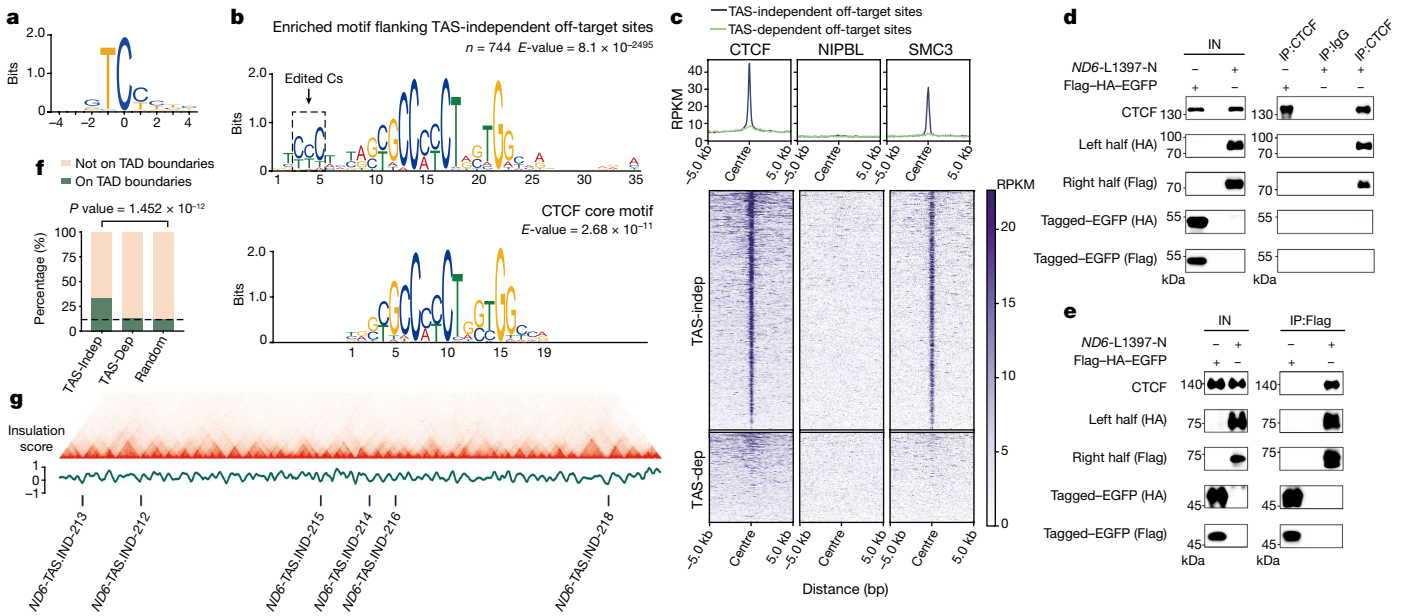


Fig. 4 | TAS-independent nDNA off-target sites are enriched at CTCF binding sites and TAD boundaries. a, Sequence logos for Cs with highest Detect-seq signal among DNA sequences at all TAS-independent off-target sites. **b**, Motif analysis of flanking sequences for the TAS-independent off-target sites using MEME discovery software (top) and with the best hit using Tomtom (bottom). **c**, Bottom, heat map showing ChIP–seq data for CTCF, NIPBL and SMC3 at the TAS-independent (indep) and TAS-dependent (dep) off-target sites. Top, the extracted line graph from the heat map; the *x*-axis represents the ±5-kb window centred on the sites and the *y*-axis shows the normalized reads per kilobase of transcript per million mapped reads (RPKM) value. **d**, Immunoblots showing the DdCBE halves co-immunoprecipitated (IP) with endogenous CTCF from *ND6*-L1397-N-transfected HEK293T cells. Images are representative of four independent biological replicates. Tagged-EGFP, EGFP tagged with Flag and HA. **e**, Immunoblots showing endogenous CTCF co-immunoprecipitated with DdCBE from *ND6*-L1397-N-transfected HEK293T cells. Images are representative of four independent biological replicates. **f**, TAS-independent off-target sites are enriched at TAD boundaries compared with random genome sampling data. *P*-value by chi-squared test. **g**, Hi-C contact data of a representative chromosome locus (chr. 7: 10,000,000–45,000,000) at 50-kb resolution. The off-target sites are indicated by vertical bars at the bottom.

(a subunit of the cohesin complex) and NIPBL[33–35] (a protein that is sufficient for cohesin loading onto DNA and loop extrusion) (Methods). We found very high levels of CTCF and cohesin but no enrichment of NIPBL at the TAS-independent off-target sites (Fig. 4c and Supplementary Fig. 21). Meanwhile, we found a low 5-methylcytosine level at these regions, consistent with the known binding preference of CTCF[36] (Supplementary Fig. 22; Methods). Taken together, these observations suggest a potential link between the TAS-independent off-target effect and the CTCF protein.

To directly assess the potential interaction between DdCBE and CTCF, we performed an in vivo co-immunoprecipitation assay for HEK293T cells transfected with *ND6*-L1397-N, *ND5.3*-L1397-C or *ND5.1*-L1397-N (Fig. 4d, e and Supplementary Fig. 23). Notably, both halves of the three DdCBEs co-immunoprecipitated with endogenous CTCF (Fig. 4d and Supplementary Fig. 23a); we also confirmed that CTCF interacts with DdCBE using the reciprocal co-immunoprecipitation (Fig. 4e and Supplementary Fig. 23b). Thus, we revealed an unanticipated physical interaction between DdCBE and CTCF.

Chromosomes are hierarchically organized into large compartments composed of smaller domains called topologically associating domains (TADs), separated by boundaries that are enriched in CTCF binding sites[37,38]. We thus analysed existing Hi-C data to further examine the potential relationship between TAS-independent off-target sites and TAD boundaries[39,40] (see Methods). Of note, about one-third ($n = 249$) of the 744 TAS-independent off-target sites co-localized with TAD boundaries; compared with randomly sampled genomic loci, TAS-independent off-target sites are significantly enriched at TAD boundaries (Fig. 4f,g). Nevertheless, we found no difference in sequence motif, Detect-seq signal strength and CTCF binding signals for off-target sites at TAD boundaries compared with those that occurred elsewhere (Supplementary Fig. 24). The mechanism for this co-localization of TAS-independent off-target sites with TAD boundaries remains to be determined.

Increasing the specificity of DdCBE

On the basis of our findings above, we propose a model for the nuclear off-target effect of DdCBE (Fig. 5a). Nuclear off-target loci can be specified by one TALE array without the limitation of matching the sequence of the other TALE repeat, and an intact deaminase can be assembled to result in TAS-dependent off-target edits. However, the intact deaminase can also be targeted to a subset of CTCF binding sites and can edit DNA at sites that are distinct from the sequence designated by the TALE arrays. Because both types of off-target sites result from DdCBE that is aberrantly localized to the nucleus, we propose that they could be prevented by inhibiting the nuclear localization of the deaminase or by inhibiting its activity in the nucleus (Extended Data Fig. 9).

We designed three different strategies to prevent off-target editing by DdCBEs (Fig. 5b): (1) we added nuclear export signal (NES) sequences to the DdCBE to reduce nuclear localization of the DdCBE protein (Extended Data Fig. 9a); (2) we simultaneously expressed $DddI_A$ (a naturally occurring inhibitor of the deaminase DddA)[1] fused to a bipartite nuclear localization signal located at both the N and C termini[1,16,41] (bis-bpNLS) to antagonize the nuclear editing activity of DdCBE (Extended Data Fig. 9b); (3) on the basis of the $DddA_{tox}$–$DddI_A$ co-crystal structure, we tested mutations of $DddA_{tox}$ that could potentially decrease its DNA binding affinity (Extended Data Fig. 9c).

We first selected eight representative nDNA off-target loci, including both TAS-dependent and TAS-independent off-target sites, for screening of the DdCBE candidates using targeted deep sequencing. Compared with the original DdCBEs, DdCBEs with a NES fused to the

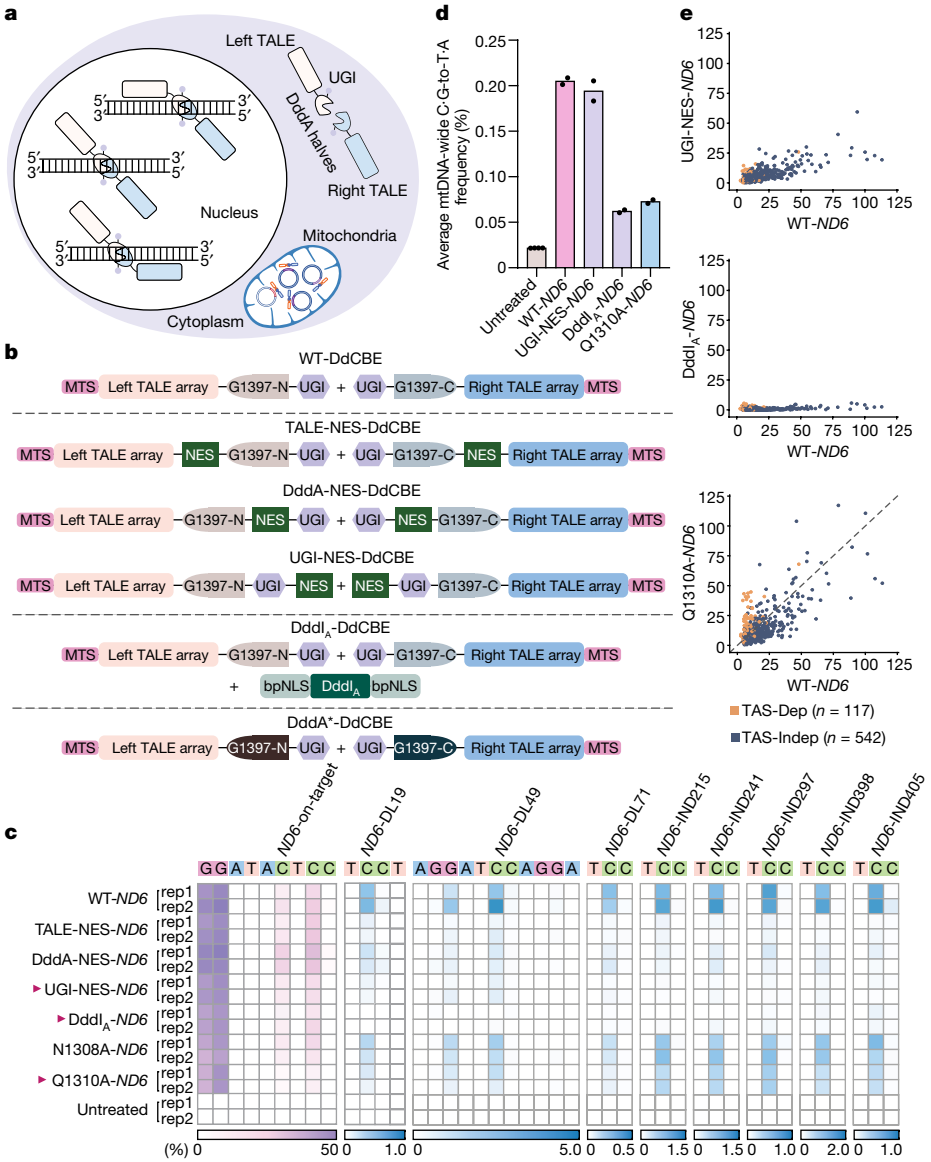


Fig. 5 | DdCBE variants with improved specificity. a, Proposed model for the off-target effects of DdCBE. **b**, DdCBE constructs based on the three different strategies to increase the specificity of DdCBE. DddA*, DddA mutants. **c**, On- and off-target editing activity of *ND6*-L1397-N variants. The ratios were assessed by targeted deep sequencing of three representative TAS-dependent and five TAS-independent nDNA off-target sites, which were induced to relatively high editing ratio by the original *ND6*-L1397-N. Red arrowheads indicate variants that were selected for further analysis. A molar ratio of DddI$_A$:ND6 of 1:1.2 was used. **d**, Average percentage of mtDNA-wide C•G-to-T•A off-target editing for the DdCBE variants and control. **e**, Comparison of Detect-seq signals at the genome-wide level between the original DdCBE and three selected variants denoted with red arrows in **c**. In **c**–**e**, *ND6* indicates the mitochondrial gene targeted by the DdCBE construct.

C terminus of TALE or uracil glycosylase inhibitor (UGI) (TALE–NES–DdCBE or UGI–NES–DdCBE) maintained on-target editing and greatly reduced off-target editing ratios from 0.222–3.58% to 0.011–0.427% or 0.010–0.499%, respectively for *ND6*, representing 8–64- or 7–101-fold lower levels of off-target editing (Fig. 5c). For *ND5.1*, the off-target sites were reduced from 0.139–2.90% to undetectable levels to 0.179% for TALE–NES–DdCBE and undetectable levels to 0.114% for UGI–NES–DdCBE (Supplementary Fig. 25). Fusing the NES to the C terminus of DddA$_{tox}$ also reduced the off-target editing but was not as effective as TALE–NES–DdCBEs or UGI–NES–DdCBEs. With proper dosage of nuclear-localized DddI$_A$ proteins, we observed mildly decreased on-target editing but a greatly reduced ratio of off-target events: when the molar ratio of DddI$_A$ to DdCBE was 1 or more, the off-target ratios were diminished to undetectable levels to 0.056% (Fig. 5c and Supplementary Fig. 26). Out of the five mutations that we selected to decrease the DNA binding affinity of DddA$_{tox}$, three showed no on-target editing activity, whereas DdCBE(N1308A) and DdCBE(Q1310A) resulted in slightly reduced on-target editing and 2- to 4-fold lower nDNA off-target editing ratios (Fig. 5c and Supplementary Fig. 27).

We carried out assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) to validate the mtDNA-wide performance of UGI–NES–DdCBE, DddI$_A$–DdCBE and DdCBE(Q1310A). We observed approximately threefold lower overall mtDNA off-target editing for DddI$_A$–DdCBE and DdCBE(Q1310A) (Fig. 5d).

Finally, we performed Detect-seq to profile the global off-target effects for the three variants with potentially improved specificity. By comparing these results with the Detect-seq results for the original DdCBE (Fig. 5e), we found that the majority of nuclear off-target editing was prevented with UGI–NES–DdCBE and DddI$_A$–DdCBE, for both TAS-dependent and TAS-independent sites. DddI$_A$–DdCBE resulted

in no off-target editing with signal strength significantly higher than the background level. However, DdCBE(Q1310A) resulted in only a small reduction in the number of genome-wide nDNA off-target sites, highlighting the necessity for genome-wide examination of off-target effects to evaluate the specificity of any altered genome editors.

Discussion

Here, using the dU-intermediate tracing method Detect-seq, we identified prevalent off-target mutations in the nuclear genome induced by the mitochondrial base editor DdCBE. The TAS-dependent off-target sites were unilateral, demonstrating that a one-sided TALE array is sufficient to guide an intact $DddA_{tox}$ to generate off-target editing events. We also detected TAS-independent nDNA off-target sites, whose genomic positions were not specified by the sequence of the TALE arrays. These TAS-independent off-target DdCBE sites differ from the Cas9-independent off-target sites of CBE in that (1) the editing ratio of TAS-independent off-target DdCBE sites is relatively high (an average of 2.21%) compared with that of Cas9-independent off-target CBE sites[26] (10^{-8} to 10^{-7} per bp) and (2) whereas Cas9-independent off-target CBE sites are random[24,25], TAS-independent off-target DdCBE sites are commonly found adjacent to CTCF binding sites. How this bacterial toxin-derived editor is recruited to a subset of CTCF binding sites in human cells remains unknown. We observed a molecular interaction between DdCBE and CTCF, although the detailed mechanism remains to be determined.

We also engineered DdCBEs with improved specificity, supporting our characterization of the off-target effects. Although more advanced DdCBE and base-editing tools are expected to emerge in future, it is important that their specificities be thoroughly evaluated. Here, we used Detect-seq as an unbiased platform to assess genome-wide specificity; the strategy of capturing editing intermediates could be used as a general approach for tracing off-target events of various genome-editing tools in both basic research and therapeutic applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04836-5.

1. Mok, B. Y. et al. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* **583**, 631–637 (2020).
2. Russell, O. M., Gorman, G. S., Lightowlers, R. N. & Turnbull, D. M. Mitochondrial diseases: hope for the future. *Cell* **181**, 168–188 (2020).
3. Vafai, S. B. & Mootha, V. K. Mitochondrial disorders as windows into an ancient organelle. *Nature* **491**, 374–383 (2012).
4. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16**, 530–542 (2015).
5. Stewart, J. B. & Chinnery, P. F. Extreme heterogeneity of human mitochondrial DNA from organelles to populations. *Nat. Rev. Genet.* **22**, 106–118 (2021).
6. Montano, V., Gruosso, F., Simoncini, C., Siciliano, G. & Mancuso, M. Clinical features of mtDNA-related syndromes in adulthood. *Arch. Biochem. Biophys.* **697**, 108689 (2021).
7. Bacman, S. R. et al. MitoTALEN reduces mutant mtDNA load and restores tRNA(Ala) levels in a mouse model of heteroplasmic mtDNA mutation. *Nat. Med.* **24**, 1696–1700 (2018).
8. Bacman, S. R., Williams, S. L., Pinto, M., Peralta, S. & Moraes, C. T. Specific elimination of mutant mitochondrial genomes in patient-derived cells by mitoTALENs. *Nat. Med.* **19**, 1111–1113 (2013).
9. Gammage, P. A., Rorbach, J., Vincent, A. I., Rebar, E. J. & Minczuk, M. Mitochondrially targeted ZFNs for selective degradation of pathogenic mitochondrial genomes bearing large-scale deletions or point mutations. *EMBO Mol. Med.* **6**, 458–466 (2014).
10. Lei, Z. et al. Detect-seq reveals out-of-protospacer editing and target-strand editing by cytosine base editors. *Nat. Methods* **18**, 643–651 (2021).
11. Zhu, C. et al. Single-cell 5-formylcytosine landscapes of mammalian early embryos and ESCs at single-base resolution. *Cell Stem Cell* **20**, 720–731.e725 (2017).
12. Shu, X. et al. Genome-wide mapping reveals that deoxyuridine is enriched in the human centromeric DNA. *Nat. Chem. Biol.* **14**, 680–687 (2018).
13. Xia, B. et al. Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **12**, 1047–1050 (2015).
14. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
15. Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, aaf8729 (2016).
16. Koblan, L. W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
17. Li, X. et al. Base editing with a Cpf1-cytidine deaminase fusion. *Nat. Biotechnol.* **36**, 324–327 (2018).
18. Wang, X. et al. Cas12a base editors induce efficient and specific editing with low DNA damage response. *Cell Rep.* **31**, 107723 (2020).
19. Sardo, L. et al. Real-time visualization of chromatin modification in isolated nuclei. *J. Cell Sci.* **130**, 2926–2940 (2017).
20. Wang, Q. et al. CoBATCH for high-throughput single-cell epigenomic profiling. *Mol. Cell* **76**, 206–216.e207 (2019).
21. Boch, J. et al. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509–1512 (2009).
22. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501 (2009).
23. Lamb, B. M., Mercer, A. C. & Barbas, C. F. III. Directed evolution of the TALE N-terminal domain for recognition of all 5′ bases. *Nucleic Acids Res.* **41**, 9779–9785 (2013).
24. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
25. Zuo, E. et al. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289–292 (2019).
26. Doman, J. L., Raguram, A., Newby, G. A. & Liu, D. R. Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. *Nat. Biotechnol.* **38**, 620–628 (2020).
27. Nakahashi, H. et al. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.* **3**, 1678–1689 (2013).
28. Schmidt, D. et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
29. Merkenschlager, M. & Nora, E. P. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu. Rev. Genomics Hum. Genet.* **17**, 17–43 (2016).
30. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
31. Shi, Z., Gao, H., Bai, X. C. & Yu, H. Cryo-EM structure of the human cohesin–NIPBL–DNA complex. *Science* **368**, 1454–1459 (2020).
32. Davidson, I. F. et al. DNA loop extrusion by human cohesin. *Science* **366**, 1338–1345 (2019).
33. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J. & Yu, H. Human cohesin compacts DNA by loop extrusion. *Science* **366**, 1345–1349 (2019).
34. Murayama, Y. & Uhlmann, F. Biochemical reconstitution of topological DNA binding by the cohesin ring. *Nature* **505**, 367–371 (2014).
35. Petela, N. J. et al. Scc2 Is a potent activator of cohesin's ATPase that promotes loading by binding Scc1 without Pds5. *Mol. Cell* **70**, 1134–1148.e1137 (2018).
36. Hashimoto, H. et al. Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. *Mol. Cell* **66**, 711–720.e713 (2017).
37. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
38. Yu, M. & Ren, B. The three-dimensional organization of mammalian genomes. *Annu. Rev. Cell Dev. Biol.* **33**, 265–289 (2017).
39. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
40. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **17**, 743–755 (2016).

Methods

Cell culture

HEK293T (ATCC, CRL-11268) and HeLa (ATCC, CCL-2) cells were cultured and maintained in DMEM (CORNING, 10-013-CVR) supplemented with 10% (v/v) FBS (Gibco), 1% (v/v) GlutaMax (Thermo Fisher Scientific) and 1% (v/v) penicillin/streptomycin (Gibco, 10378016) at 37 °C with 5% CO_2. The subculture of cells was performed every 2 days and only passages 4–6 were used for subsequent experiments. All cells were routinely tested for mycoplasma contamination with a mmycoplasma detection kit (TransGene Biotech, FM311-01).

Plasmid cloning

PCR was performed using TransStart FastPfu DNA Polymerase (TransGene Biotech, AP221-01). Most plasmids were constructed by Gibson assembly using NEBuilder HiFi DNA Assembly Master Mix (NEB) and transformed into Trans1-T1 chemically competent cells (TransGene Biotech). For construction of the original DdCBE plasmids, MitoTALE genes were assembled through the Advanced Ulti-MATE system[42] and constructed into pGL3-TALEN vector; genes encoding MTS, $DddA_{tox}$ and UGI were synthesized as gene blocks and codon optimized for mammalian expression (Rui Biotech); the gene fragments encoding mitoTALE, MTS, $DddA_{tox}$ splits and UGI were respectively amplified and cloned into the pCMV plasmid backbone by Gibson assembly. Then deletion plasmids lacking either one or two TALE arrays were constructed based on the original DdCBE plasmids. For construction of plasmids to improve DdCBE specificity, NES sequences were incorporated into three different positions of the original DdCBE plasmids by Gibson assembly (Fig. 5b); the $DddI_A$ gene was codon optimized, synthesized and inserted into a pCMV backbone with bis-bpNLS; and the DdCBE variants with mutated DddA were generated by QuickChange site-directed mutagenesis based on the original DdCBE constructs.

Transfections

For Detect-seq and verification by targeted deep sequencing, HEK293T cells were seeded on 6-well cell culture plates (NEST Biotechnology) at a density of 3.2×10^5 cells per ml (2 ml total per well). After 18–22 h, transfection was performed at a cell density of approximately 60%. Cells in each well were transfected with 3,500 ng of each DdCBE monomer, using 21 µl Lipofectamine LTX and 7 µl PLUS reagent (Thermo Fisher Scientific). For screening of DdCBE variants with improved specificity by targeted deep sequencing, HEK293T cells were seeded in 24-well cell culture plates at a density of 3.2×10^5 cells per ml (0.5 ml total per well). Cells were transfected with 840 ng of each DdCBE monomer, using 5.04 µl Lipofectamine LTX and 1.68 µl PLUS reagent. As in the case of $DddI_A$ co-expression, the $DddI_A$ plasmid was co-transfected according to the molar ratio to DdCBE monomer. Taking $DddI_A$-*ND6* (1:1) as an example, 840 ng of each *ND6*-DdCBE monomer (~5.8 kb) and 680 ng $DddI_A$ (~4.7 kb) were simultaneously transfected to cells in 24-well plates using 7.08 µl Lipofectamine LTX and 2.36 µl PLUS reagent. Cells were then collected after 72 h of transfection. Genomic DNA was freshly extracted using the CWBIO universal genomic DNA kit (CWbiotech, CW2298M) and stored in EB buffer (10 mM Tris-HCl, pH 8.0) at −80 °C.

ATAC-seq for mitochondrial genome sequencing

ATAC-seq was performed as previously reported[1,43]. Cells cultured in 24-well plates were trypsinized and washed with cold PBS; then cells were counted using a cell counting chamber. Cells (~10^4) were pelleted (500 RCF at 4 °C for 5 min) and lysed in 50 µl cold and freshly prepared lysis buffer (10 mM Tris-HCl, 10 Mm NaCl, 3 mM $MgCl_2$ and 0.1% (v/v) NP-40). Lysates were incubated on ice for 3 min. Then they were pelleted (500*g* at 4 °C for 5 min) and tagmented with 2.5 µl self-assembled Tn5 transposase in a 20 µl reaction system containing 1× TD buffer (20 mM Tris-HCl PH 7.6, 10 mM $MgCl_2$, 20% (v/v) dimethyl formamide, 0.1% (v/v) NP-40 and 0.3× PBS). Samples were incubated at 37 °C for 30 min on a thermomixer at 300 rpm. The tagmented DNA was purified using the DNA Clean and Concentrator-5 Kit (Vistech) and eluted in 20 µl ultrapure water. All 20 µl of eluate was amplified using universal primer (1 µM), indexed primers (1 µM) and 2× NEBNext Q5 Hot Start HiFi PCR Master Mix (NEB) in a total volume of 50 µl, using the following procedure: 72 °C for 5 min, 98 °C for 30 s, then 12 cycles of (98 °C for 10 s, 63 °C for 30 s, and 72 °C for 60 s), followed by a final 72 °C extension for 2 min. The final library was purified and size-selected using Agencourt AMPure XP beads (Beckman Coulter). After qualification using Qubit dsDNA HS Assay kit (Invitrogen) and Agilent 4150 TapeStation System, all libraries were sequenced on Illumina HiSeq X Ten.

Detect-seq

As shown in Extended Data Fig. 1, extracted genomic DNA was sheared into fragments of approximately 300 bp length using a Covaris focused ultrasonicator instrument (ME220). End repair was performed on 5 µg DNA fragments and 10 pg spike-in model sequences using NEBNext End Repair Module (NEB, E6050) with *Escherichia coli* ligase (NEB, M0205) added to repair nicks in DNA. Then endogenous 5-formyl-2′-deoxycytidines (5fdCs) were protected by 10 mM *O*-ethylhydroxylamine (Aldrich, 274992) in 100 mM MES buffer (pH 5.0) at 37 °C for 6 h. dA-tailing was performed by NEBNext dA-Tailing Module (NEB, E6053).

Damage repair was used to remove potential signal noise that may interfere the subsequent labelling, such as abasic sites, single-stranded breaks, nicks and others. Specifically, DNA was incubated with 1 µl dNTP (2.5 mM each), 1 µl NAD^+ (NEB, B9007), 5 µl 10× NEBuffer 3, 2 µl Endo IV (NEB, M0304), 1 µl Bst full-length polymerase (NEB, M0328) and 2 µl Taq ligase (NEB, M0208) in a total volume of 50 µl for 1 h at 37 °C and 1 h at 45 °C. The products were purified using 2× Agencourt AMPure XP beads and then subjected to in vitro BER labelling reaction using 200 nM biotin-dUTP, 800 nM 5fdCTP, 200 nM dATP, 200 nM dGTP in 1× NEBuffer 3, 1 µl NAD^+, 1 µl UDG (NEB, M0280), 1.5 µl Endo IV, 0.8 µl Bst full-length polymerase, 1.7 µl Taq ligase to a total volume of 50 µl for 40 min at 37 °C. The product was purified using 2× Agencourt AMPure XP beads and then incubated in 10 mM Tris-HCl (pH 7.0) containing 75 mM of malononitrile at 37 °C for 20 h on a thermomixer (Eppendorf) at 850 rpm.

Each sample of labelled fragments was enriched using 10 µl streptavidin C1 beads (Invitrogen) according to the manufacturer's instruction. Ligation of Y adaptors (NEBNext Quick Ligation Module, E6056) to the DNA was performed on streptavidin C1 beads, followed by three washes with 1× B&W buffer (5 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween-20) to remove free adaptors. Treatment with 150 mM NaOH was performed to remove the complementary chain. The DNA library on C1 beads was eluted in nuclease-free water after heating at 95 °C for 3 min. All of the eluate was amplified using MightyAmp DNA Polymerase (NEB) for 2 cycles followed by amplification using Q5 Hot Start HiFi PCR Master Mix (NEB) for 8 or 9 cycles. The final library was purified using 0.9× Agencourt AMPure XP beads and subjected to quantification using the Qubit dsDNA HS assay kit (Invitrogen) and fragment analyzer. To evaluate the efficiency and specificity of each batch of Detect-seq experiments, quantitative PCR (qPCR) and Sanger sequencing were performed as previously described on spike-in molecules of each sample[10]. All libraries were finally sequenced using Illumina Hiseq X Ten and MGISEQ-2000.

Immunofluorescence staining for unfixed nuclei

HeLa cells transfected with different DdCBE constructs (expressing HA-tagged left half and Flag-tagged right half) or vector plasmids were harvested by trypsinization. Then nuclei were isolated as described[44]. Approximately 2.5×10^5 nuclei for each sample were stained 1 h on ice with 50 µl 5% fetal bovine serum in 1× phosphate-buffered saline (5% FBS/PBS) solution containing diluted primary antibodies (anti-HA (Abcam, ab9110, 1:200); anti-Flag (Sigma-Aldrich, F1804, 1:100)). Nuclei

were washed twice in 5% FBS/PBS and stained for 1 h on ice with 100 µl 5% FBS/PBS containing diluted secondary antibodies (Alexa Fluor conjugated anti-rabbit (HA tag) or anti-mouse (Flag tag), Thermo Fisher, 1:500). One hundred microlitres of 5% FBS/PBS with 2 ul DAPI (to a final concentration of 10 µg ml^{-1}) was added and mixed thoroughly. The nuclei were stained for another 30 min on ice. The nuclear pellet was washed twice with 5% FBS/PBS, resuspended in 50 µl glycerol storage buffer (50 mM Tris-HCl (pH 8.0), 2 mM $MgCl_2$, 0.1 mM EDTA, 30% glycerol) and used for imaging experiments. *z*-stack images for these unfixed nuclei were collected at 0.4 µm intervals under the same exposure condition by DeltaVision OMX SR. Then, using Imaris9.7, the 3D mean fluorescence intensity per voxel was calculated for each scanned nucleus after surface modelling based on DAPI (two examples are shown in Supplementary Videos 1 and 2).

Cell fractionation followed by western blot and fixation-based immunofluorescence

Subcellular fractions were prepared using a nuclear/cytosol fractionation kit (BioVision, K266-100) and fluorescence-activated cell sorting (FACS) steps. Seventy-two hours after transfection, HEK293T cells were labelled with 100 nM MitoTracker Deep Red (Thermo, M22462) for 30 min at 37 °C in a 5% CO_2 incubator. Cells (6×10^6) were collected by centrifugation at 600*g* for 5 min at 4 °C and resuspended in 600 µl CEB-A mix containing DTT and protease inhibitors, followed by vortexing for 15 s and incubation on ice for 10 min. After that, 33 µl of ice-cold CEB-B were added to the lysates, followed by vortexing for 5 s, incubation on ice for 1 min and vortexing for another 5 s. The lysates were centrifuged at 16,000*g* for 5 min at 4 °C. The supernatants (cytoplasmic extract) were immediately transferred to a clean pre-chilled tube and saved as the cytoplasmic extract fraction. The pellets containing the nuclei were washed twice with 600 µl cold PBS, followed by incubation in 300 µl PBS containing 10 µg ml^{-1} DAPI for 10 min. Then the nuclei were subjected to FACS sorting via DAPI signal and collected in PBS containing 2% FBS. The glow-cytometric pseudo-colour plots were processed using BD FACSDiva (Version 8.0.1) and FlowJo (Version 10.0.7r2).

For western blot analysis, ~1×10^6 sorted nuclei were centrifuged and resuspended in 100 µl of ice-cold nuclear extraction buffer mix, followed by vortexing for 15 s and incubation on ice for 10 min. This process of vortexing and incubation was repeated 4 times. Then the mixtures were centrifuged for 10 min at 16,000*g* at 4 °C. The supernatants were immediately transferred to a clean pre-chilled tube and saved as the nuclear extract fraction, while the pellets containing the chromatin and chromatin-bound proteins were resuspended in 125 µl 1× SDS loading buffer. Chromatin fractions, nuclear extract fractions and cytoplasmic extract fractions from ~2×10^3 cells were analysed by 15% SDS–PAGE for western blot (Extended Data Fig. 5a and Supplementary Fig. 1). Anti-ATP5a (Abcam, ab14748, 1:5,000), anti-GAPDH (Abcam, ab8245, 1:2,000) and anti-H3 (EASYBIO, BE3015, 1:10,000) antibodies were used to indicate the results of cell fractionation; anti-HA (Abcam, ab1424, 1:5,000) and anti-Flag (Sigma-Aldrich, F1804, 1:2,000) antibodies were used to show the localization of the left half and right half of DdCBE, respectively; HRP-conjugated goat anti-mouse IgG (CWBIO, CW0102, 1:5,000) was used as the secondary antibody.

For fixation-based immunofluorescence experiments, ~5×10^4 FACS-sorted nuclei were centrifuged onto polylysine-coated microscope adhesion slides (Thermo, P4981) and fixed in 4% paraformaldehyde/PBS for 15 min at room temperature. The slides were washed three times with PBS and the nuclei on the slides were permeabilized with 0.5% Triton X-100/PBS for 15 min at room temperature. After three washes with PBS, the nuclei were blocked in 5% BSA/PBS for 1 h at room temperature. Then the nuclei were incubated with primary antibody (anti-HA (Abcam, ab9110, 1:100) and anti-Flag (Sigma-Aldrich, F1804, 1:100)) overnight at 4 °C, followed by three washes with PBS. The nuclei were incubated with Alexa Fluor 568 goat anti-rabbit IgG (Thermo, A-11036, 1:100), Alexa Fluor 488 goat anti-mouse IgG (Proteintech, SA00006-1, 1:100) and 10 µg ml^{-1} DAPI for 1 h at room temperature in dark chamber. The slides were washed three times with PBS, and then the nuclei were mounted on slides using SlowFade Diamond Antifade Mountant (Thermo, S36968) for imaging. Images were obtained using a 60× oil objective with the Nikon A1R confocal laser scanning microscope system. Acquired images were then processed using Fiji (version 2.1.0).

Targeted deep sequencing

Primers containing the paired Illumina adaptor sequences in the overhangs were designed based on regions flanking the off-target sites (Supplementary Table 2). A 10-nt barcode was also added into each primer pair[10] to reduce the detection limit from ~0.1% to ~0.005%. Genomic DNA (10–100 ng) was used for the first round of PCR amplification using NEBNext Q5U Hot Start HiFi PCR Master Mix (NEB, M0515L) for approximately 10 cycles. Q5U is capable of reading and amplifying templates containing uracil bases and hence ensures accurate measurement of off-target editing level. The PCR products were purified with 1× Agencourt AMPure XP beads and eluted in nuclease-free water. The second round of amplification was performed on the purified DNA samples with different index primers for about 15 cycles. The PCR products were purified with 0.8× Agencourt AMPure XP beads and eluted in nuclease-free water. Targeted deep sequencing for on-target sites in the mtDNA was performed as previously described[1]. The libraries were quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a Qubit 2.0 fluorometer (Invitrogen), and then pooled together for high-throughput sequencing by Illumina HiSeq X Ten or MGISEQ-2000.

In situ ChIP–seq

Low-input in situ ChIP was performed as previously described with minor modifications[20]. Briefly, transfected HEK293T cells were harvested and cross-linked with 0.25% formaldehyde on ice for 5 min. After fixation, cells were quenched by adding 2.5 M glycine and incubated on ice for 5 min. The fixed cells were washed twice with 1% BSA/PBS and conjugated with adequate prepared Con-A beads. Next, the cells–beads mixture was incubated with the primary antibody (anti-Flag, Sigma-Aldrich, F1804, 1:100), the secondary antibody (donkey anti-mouse-Alexa 488, Invitrogen, A21202, 1:500), and purified PAT-MEA/B, step by step. After washing out free PAT-MEA/B, tagmentation was performed at 30 °C for 1 h in a thermal cycler and stopped by adding 10 µL 40 mM EDTA. Cells were then lysed at 55 °C for at least 3 h in lysis buffer. For library preparation, DNA fragments in the same tube were enriched via 10–18 cycles of PCR amplification with Nextera i5 index primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC-3′) and Nextera i7 index primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′). After PCR, the library was purified and selected for 200- to 1,000-bp fragments for sequencing. The libraries were sequenced with Illumina Nova-Seq 6000 sequencer.

Co-immunoprecipitation assay

Around 10^7 HEK293T cells transfected with DdCBE (expressing HA-tagged left half and Flag-tagged right half) or intact DddA (expressing bpNLS-linked, inactivated DddA-ugi tagged with both HA and Flag) or control (expressing HA–Flag-tagged EGFP) plasmids were collected and pelleted by centrifugation at 1,000*g* for 5 min. The cell pellet was washed three times with ice-cold PBS and finally resuspended with 300 µl ice-cold lysis buffer (50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 1 mM EDTA, 1% TritonX-100, 0.1% SDS, 1 mM NaF, 1 mM PMSF and 1/50 protease inhibitor cocktail (Roche)) and incubated on ice for 30 min. Then the cell lysates were immediately diluted by adding 400 µl ice-cold IP-wash buffer (20 mM Tris-HCl (pH 7.4), 100 mM NaCl, 1 mM EDTA, 0.1% NP-40, 1 mM PMSF, and 1:100 protease inhibitor cocktail (Roche)) and centrifuged at 14,800 rpm for 15 min at 4 °C to remove cell debris. 30 µl of the resultant supernatant was saved as input.

For co-immunoprecipitation with endogenous CTCF, the rest of the supernatant was pre-cleared with 30 µl prepared protein A beads (Invitrogen, 10001D) for 2–3 h at 4 °C on a rotator. The beads were removed with a magnetic stand, and the supernatant was then incubated with 2 µg anti-CTCF antibody (Abcam, ab128873) or normal rabbit IgG (Biodragon, BF01001) at 4 °C overnight. The reaction mixture was incubated with 60 µl prepared protein A beads at 4 °C for another 2 h. The immunoprecipitated complex was washed eight times with 1 ml ice-cold IP-wash buffer and finally eluted in 2× SDS loading buffer at 95 °C for 5 min. The eluted products were saved as the immunoprecipitate.

For co-immunoprecipitation with DdCBE, the rest of the supernatant was incubated with 25 µl prepared anti-Flag M2 magnetic beads (Millipore, M8823) overnight at 4 °C. The beads were collected by magnetic stand and subsequently washed eight times with 1 ml ice-cold IP-wash buffer, followed by incubation with the ice-cold elution solution (0.4 mg ml^{-1} 3× Flag peptide (Sigma-Aldrich) in IP-wash buffer) at 4 °C for 2 h. The eluted products were saved as the immunoprecipitate.

All input and immunoprecipitation products were analysed by 8% SDS PAGE for western blotting with anti-CTCF (Abcam, ab128873, 1:2,000), anti-HA (Abcam, ab9110, 1:1000; Abcam, ab1424, 1:2,000), anti-Flag (Sigma-Aldrich, F1804, 1:2,000; ABclone, AE063, 1:1,000) antibodies, HRP-conjugated goat anti-mouse IgG (CWBIO, CW0102, 1:5,000) and HRP-conjugated goat anti-rabbit IgG (CWBIO, CW0103, 1:5,000).

Detect-seq mapping

We processed Detect-seq data as our previous paper described[10]. In brief, we removed adapter sequences from raw sequencing reads by cutadapt software (version 1.18)[45]. We mapped those reads to the reference genome (hg38) with the sequence-converted aligner Bismark (version 0.22.3)[46] using default settings. We collected the unmapped reads and the reads with MAPQ less than 20, then re-mapped those reads with BWA MEM (version 0.7.17)[47] and GATK IndelRealigner (v.3.8.1)[48] using default parameters. The Bismark- and BWA-generated BAM files were merged and sorted by reference coordinate with samtools sort command (version 1.9)[49]. PCR duplications were removed from the sorted BAM files by Picard MarkDuplicates (version 2.0.1)[50].

Mutation reads and count normalization

We considered sequencing reads with no less than 2 tandem C-to-T mutations as Detect-seq mutation reads and sequencing reads without C-to-T mutation as non-mutation reads. According to this definition, we calculated the normalized mutation reads count for each Detect-seq signal region using the following formula:

$$\text{Normalized mutation reads count} = \frac{\text{Region mutation reads count}}{\text{Total mapped reads count}/10^6} \times 100$$

Identification of significantly enriched Detect-seq signal regions

To search genome-wide tandem C-to-T signals, we first converted BAM files to mpileup files using the samtools mpileup command (version 1.9)[49] with the parameters -q 20 -Q 20. Then we generated .bmat and .pmat files from those mpileup files by Detect-seq tools parse-mpileup and bmat2pmat commands with default settings. We next searched the genome-wide tandem C-to-T signals using the pmat-merge command. Those tandem C-to-T signals were filtered with mpmat-select command with settings -m 4 -c 6 -r 0.01–RegionPassNum 1 –RegionToleranceNum 3. Then we used find-significant-mpmat to perform the statistical test for each filtered region. In brief, a Poisson one-sided test was performed; the parameter λ in this test was set to the normalized Detect-seq mutation reads count in the control sample. After the statistical test, the *P*-value was adjusted with the Benjamini and Hochberg method to control the false discovery rate. All scripts used in this step were collected into the Detect-seq tools.

Alignment for pTBSs

To find a putative binding site for TALE (pTBS), we extracted sequences from the reference genome hg38 and aligned them with the TALE arrays designed binding sequence using a semi-global alignment algorithm. The alignment with highest score was reported as the putative TALE array binding site. Considering that the repeat-variable diresidue NN could recognize both G and A, we set A:G mismatch alignment score as +3, the other mismatch alignment score as −4, match score as +5, gap open score as −24, and gap extension score as −8.

Identification of the TAS-dependent and TAS-independent off-target sites

We identified DdCBE off-target sites by comparing the Detect-seq signals between GFP samples and DdCBE-treated samples. Any region complying with the following criterion was considered a DdCBE off-target site: false discovery rate less than 0.01; fold change of normalized mutation reads count in the DdCBE-treated sample to normalized mutation reads count in the GFP sample larger than 2; mutation reads count in the GFP sample no larger than 1, and mutation reads count in the DdCBE-treated sample no less than 10. The identified DdCBE off-target sites with normalized mutation signals not responding to TALE deletion were considered TAS-independent off-target sites. The remaining DdCBE off-target sites showing no higher mutation signal than background level after deletion of any TALE part (normalized signal no more than 1) were considered TAS-dependent off-target sites. A small portion of unclassified off-target sites were added as extended lists in Supplementary Table 1.

In situ ChIP–seq analysis

We analysed DdCBE in situ ChIP–seq data as previously described[20]. More specifically, we used cutadapt (version 1.18)[45] to remove sequencing adapters and mapped clean reads to reference genome hg38 with Bowtie2 (version 2.4.2)[51]. The additional settings "–no-mixed –no-unal –no-discordant –dovetail –very-sensitive-local -X 2000" were used for fast and sensitive reads alignment. Next, we used Picard (version 2.0.1)[50] to remove PCR duplication and the samtools (version 1.9)[49] view command to select alignments with MAPQ over 20. Then we used MACS2 (version 2.1.0)[52] to identify enriched peaks with default settings. Finally, peaks with *q*-value smaller than 0.01 and enrichment larger than 5-fold were considered for downstream analysis. The correlation heat map plots were generated by deepTools (version 3.1.3)[53] bamCoverage and plotHeatmap programs with "–normalizeUsing RPKM" settings. The intersection analysis of peaks was performed with Bedtools (version 2.27.1).

Targeted deep sequencing data analysis

The raw reads (FASTQ) of targeted deep sequencing were grouped by the unique molecular identifier (UMI). UMI groups contained less than three reads were discarded. We considered the most frequent reads in the same UMI groups as the consensus reads. Then we used cutadapt (version 1.18)[45] to remove adapter sequences from the consensus reads. Cleaned reads were mapped to the targeted loci using BWA MEM (version 0.7.17)[47] with default parameters. Then the BAM files were used to generate mutation information in .mpileup format using the samtools (version 1.9)[49] mpileup command with parameters -q 20 -Q 20. Finally, the .mpileup files were converted to .bmat files using the Detect-seq tools parse-mpileup commands with default settings.

ATAC-seq data analysis

First we used cutadapt (version 1.18)[45] to remove adapter sequences from mitochondrial ATAC-seq sequencing data. Then we mapped the cleaned reads to the human mitochondrial genome (extracted from reference hg38) using the Bowtie2 aligner (version 2.4.2)[51] with default settings. After the mapping step, we used the samtools (version 1.9)[49] view command with parameters -hb -q 30 -F 4 -F 8 to

select high-quality alignments. Next, we used Picard (version 2.0.1)[50] to remove PCR duplications. The BAM files were used to generated mutation information in .mpileup format using the samtools (version 1.9)[49] mpileup command with parameters -q 30 -Q 30. Finally, we used the Varscan2 (version 2.4.4)[54] mpileup2snp command to identify the potential mutations.

Hi-C data analysis

The Hi-C data used in this study were downloaded from the GEO database under accession number GSE44267. Those sequencing reads were mapped by HiC-Pro (version 3.00)[55] to the hg38 reference genome. Then the valid Hi-C interactions were collected and normalized with the KR method by HiCExplorer (version 3.6)[56]. The insulation score and TAD boundaries were calculated using the HiCExplorer hicFindTADs command with a 25 kb resolution Hi-C normalized matrix.

Public data analysis

All sequencing reads from public data were processed with the ENCODE Data Standards and Prototype Processing Pipeline (https://www.encodeproject.org/data-standards/).

Statistical analysis

The chi-squared test, Student's *t*-test and Pearson's correlations in this study were performed in the R environment (version 3.6).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data generated for this paper have been deposited at NCBI Gene Expression Omnibus (GEO) and are available under GEO accession number GSE173859 (Detect-seq data), GSE173689 (ATAC-seq data and in situ ChIP–seq data) and GSE176089 (targeted deep sequencing data). hg38 was used as the reference genome. The Hi-C, DNase-seq, Bisulfite-seq and ChIP–seq data were downloaded from the GEO or ENCODE database; accession numbers of these public data sets are available in Supplementary Table 5.

Code availability

Detect-seq tools, including several Python scripts, were deposited on GitHub (https://github.com/menghaowei/Detect-seq). Detect-seq tools can help to perform Detect-seq analysis, including but not limited to Detect-seq signal finding, enrichment testing, off-target sites identification, TALE sequence alignment and alignment results visualization.

41. Suzuki, K. et al. In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature* **540**, 144–149 (2016).
42. Yang, J. et al. ULtiMATE system for rapid assembly of customized TAL effectors. *PLoS ONE* **8**, e75649 (2013).
43. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21–29 (2015).
44. Neely, A. E. & Bao, X. Nuclei isolation staining (NIS) method for imaging chromatin-associated proteins in difficult cell types. *Curr. Protoc. Cell Biol.* **84**, e94 (2019).
45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* https://doi.org/10.14806/ej.17.1.200 (2011).
46. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
49. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
50. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol* **9**, R137 (2008).
53. Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
54. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
55. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
56. Wolff, J. et al. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).

Acknowledgements We thank W. Wei for providing related plasmids; H. Cheng for sharing the antibodies for mitochondrial markers; W. Xie and X. Lu for discussion about the ChIP assay; X. Zhang and Chuyun Shao for help with ATAC-seq experiments and data processing; National Center for Protein Sciences at Peking University for assistance with FACS, imaging, sequencing, Imaris, Fragment Analyzer and Agilent 4150 TapeStation System; C. Shan, L. Fu, S. Qin and Y. Guo for assistance with immunofluorescence experiments, FACS and image processing; and G. Li and X. Zhang for assistance with NGS experiments. Bioinformatics analysis was performed on the High-Performance Computing Platform of the School of Life Sciences and High-Performance Computing Platform of the Center for Life Science. This work was supported by the National Natural Science Foundation of China (nos. 21825701, 91953201, 92153303 and 22107006), National Key R&D Program (2019YFA0110900 and 2019YFA0802200) and China Postdoctoral Science Foundation (2020M680218, 2021M700238). L. Liu was supported in part by the Postdoctoral Fellowship of Peking-Tsinghua Center for Life Sciences.

Author contributions Z.L., H.M. and C.Y. conceived and led the project. Z.L., H.M., L.L. and C.Y. designed the experiments to investigate the DdCBE off-target effect, which were performed by Z.L. and L.L. H.M. analysed Detect-seq, in situ ChIP–seq and the downloaded public data. H.Z. analysed ATAC-seq and targeted deep sequencing data. Z.L. performed the co-immunoprecipitation and non-fixation immunofluorescence assays. L.L. conducted the cell fractionation assay and the sequential western blot and fixation-based immunofluorescence experiments. X. R. and Y. Y. assisted the experiments and data processing. Z.L. and C.Y. designed the in situ ChIP experiments with the advice of A.H. M.L. performed the in situ ChIP–seq experiments with cell samples prepared by Z.L. Z.L., H.M., L.L., H.Z. and C.Y. wrote the paper with the help of X. R. and H.W.

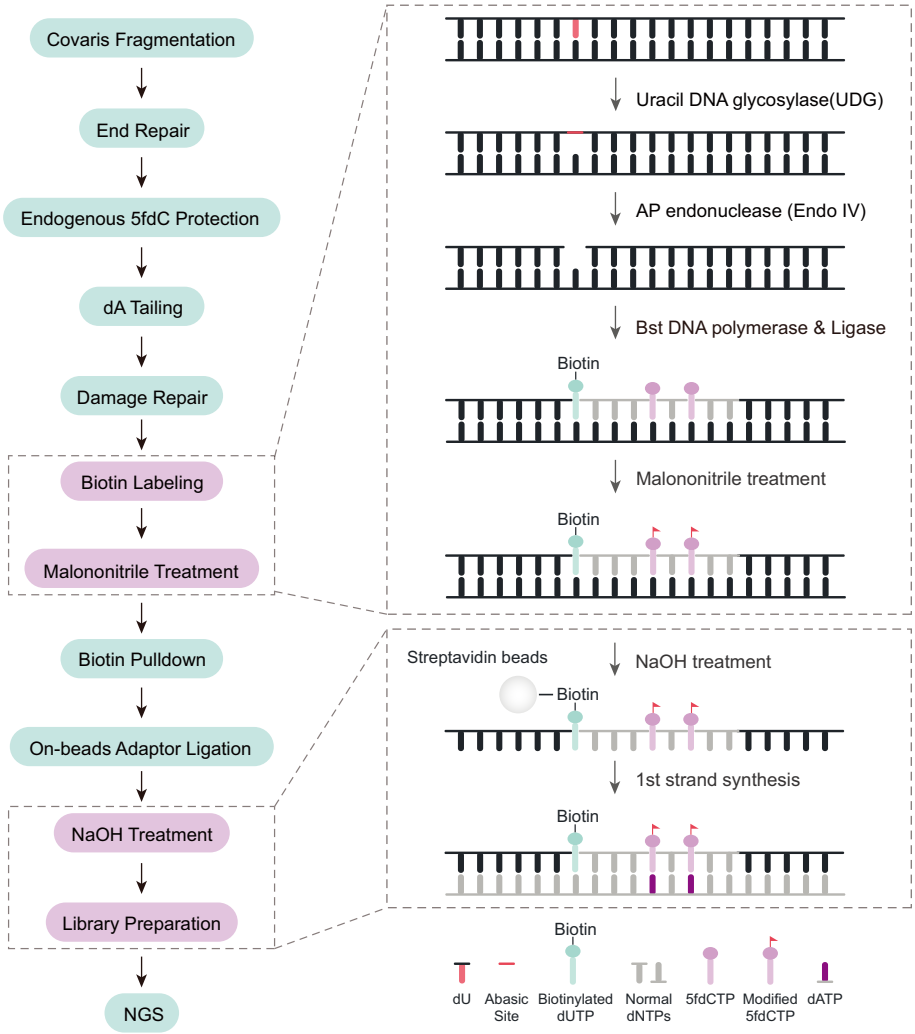
Competing interests Peking University has filed patent applications on Detect-seq and optimized DdCBE variants described in this study, listing Z.L., H.M., Z.C.L., L.L, H.Z. and C.Y. as inventors.

Additional information

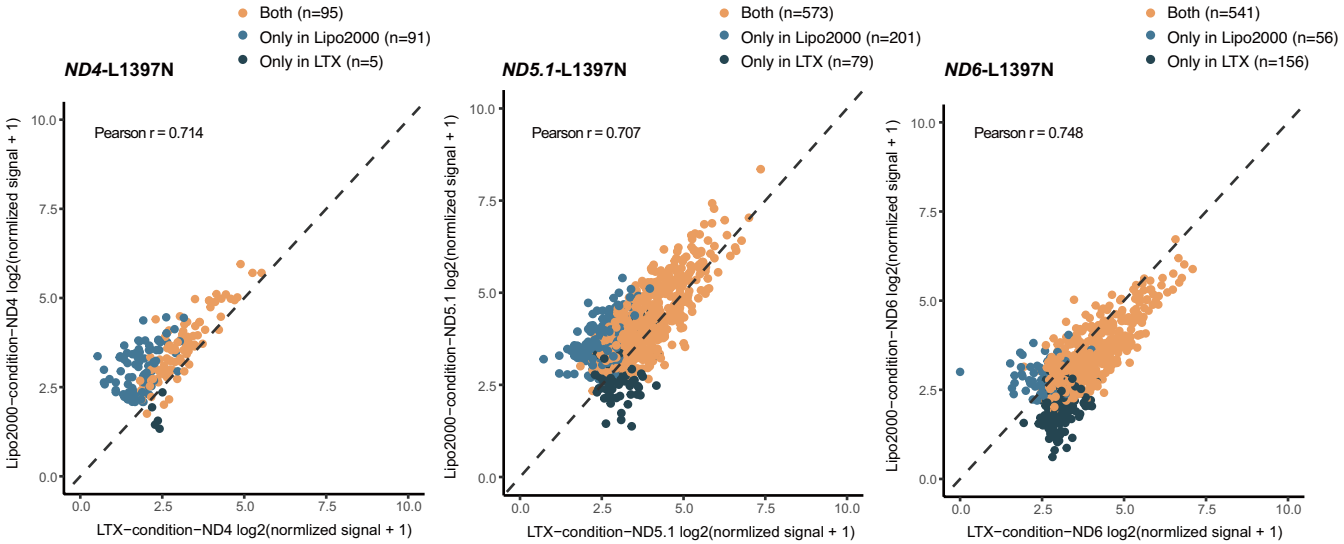
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04836-5.

Correspondence and requests for materials should be addressed to Chengqi Yi.

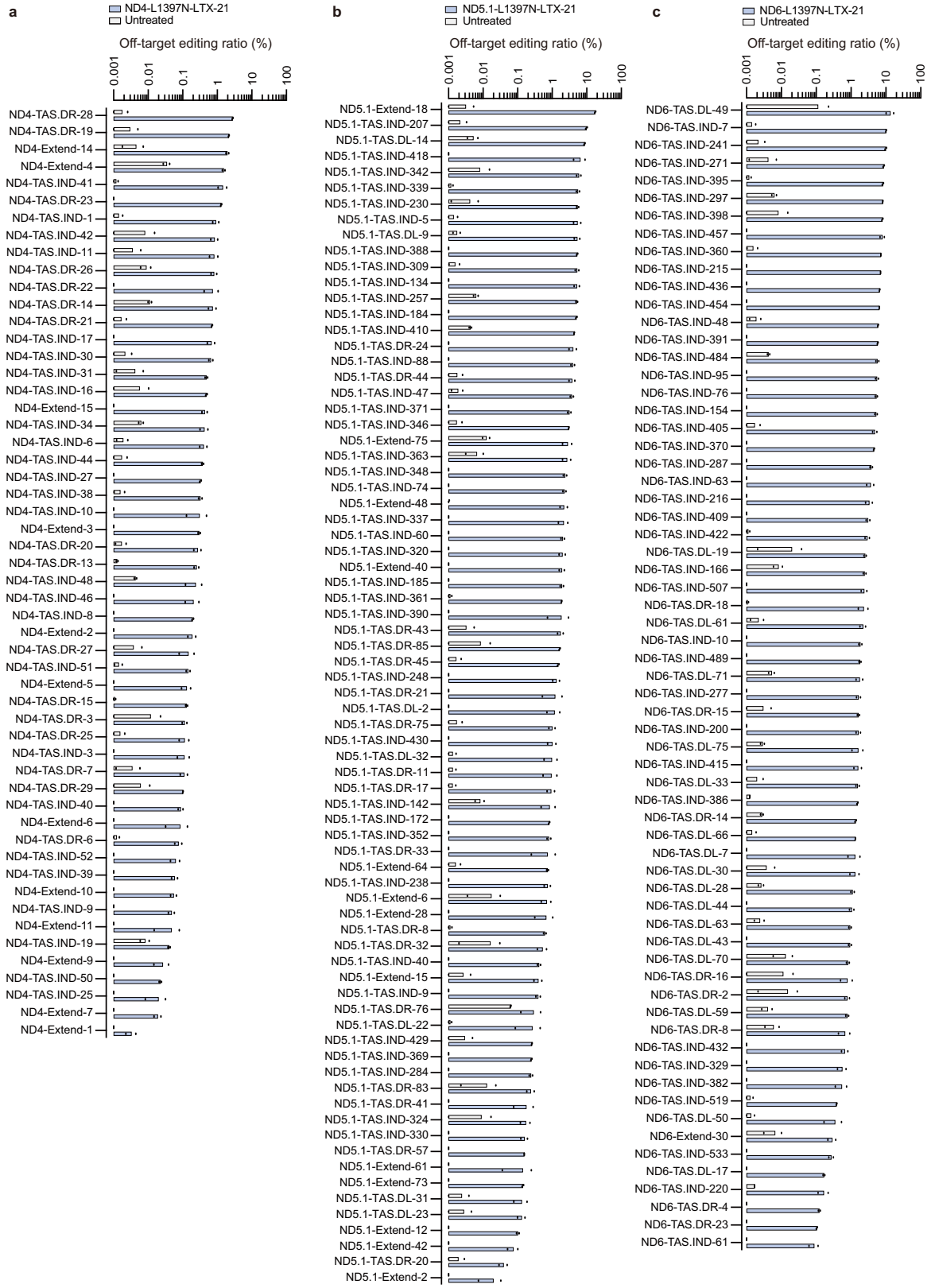
Peer review information *Nature* thanks Bryan Dickinson and Fyodor Urnov for their contribution to the peer review of this work.



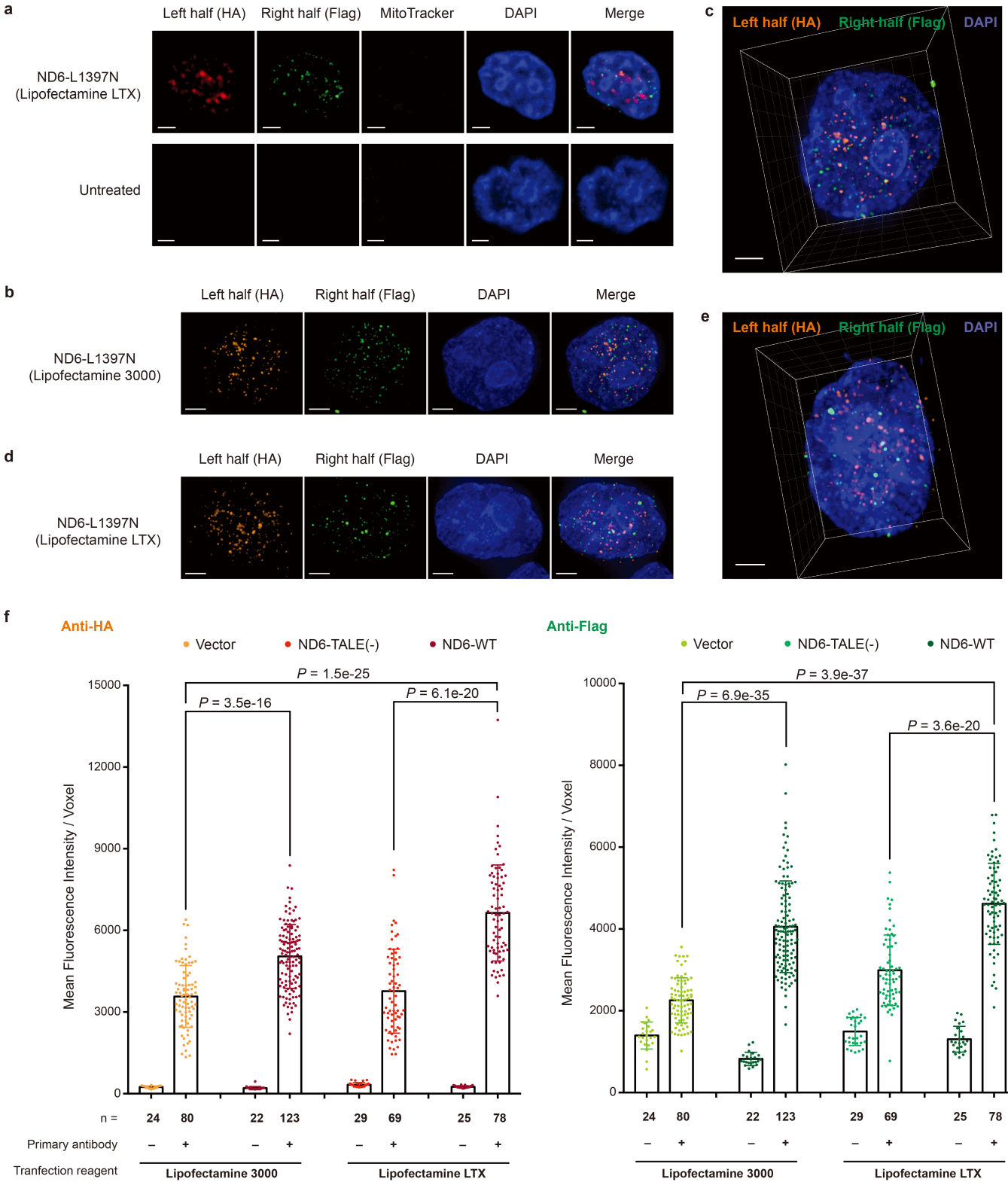
Extended Data Fig. 1 | Workflow of Detect-seq. Endogenous 5fdC was protected by *O*-ethylhydroxylamine ($EtONH_2$). Damage repair step eliminates endogenous DNA damages including abasic sites (AP), single strand breaks (SSB), etc. Deoxyuridine (dU) generated by DdCBE in vivo was labeled by the in vitro reconstituted base excision repair (BER) reaction: UDG specifically recognizes and cleaves dU, leaving abasic sites; Endo IV removes abasic sites, leaving 3′-OH remnant; *Bst* DNA polymerase initiates DNA strand replacement after the 3′-OH; ligase sews the final nicks. Through the so-called "nick translation" activity of *Bst* polymerase during this step, biotinylated dUTPs and 5fdCTPs were incorporated 3′ to dU. Malononitrile treatment marks the incorporated 5fdCs, generating a characteristic tandem C-to-T mutation pattern to trace DdCBE edits. Biotin pulldown followed by NaOH treatment enriches DdCBE edited DNA fragments and enhances Detect-seq signals.



Extended Data Fig. 2 | Comparisons of Detect-seq signals for off-target edits under two different transfection conditions. The two conditions are: $4x10^5$ seeded HEK293T cells on 6-well plates were transfected with 4 μg of each monomer using 12 μl Lipofectamine 2000; or, $6.4x10^5$ cells were transfected with 3.5 μg of each monomer using 21 μl Lipofectamine LTX. The Detect-seq signals are highly consistent between the two conditions for all three DdCBEs.

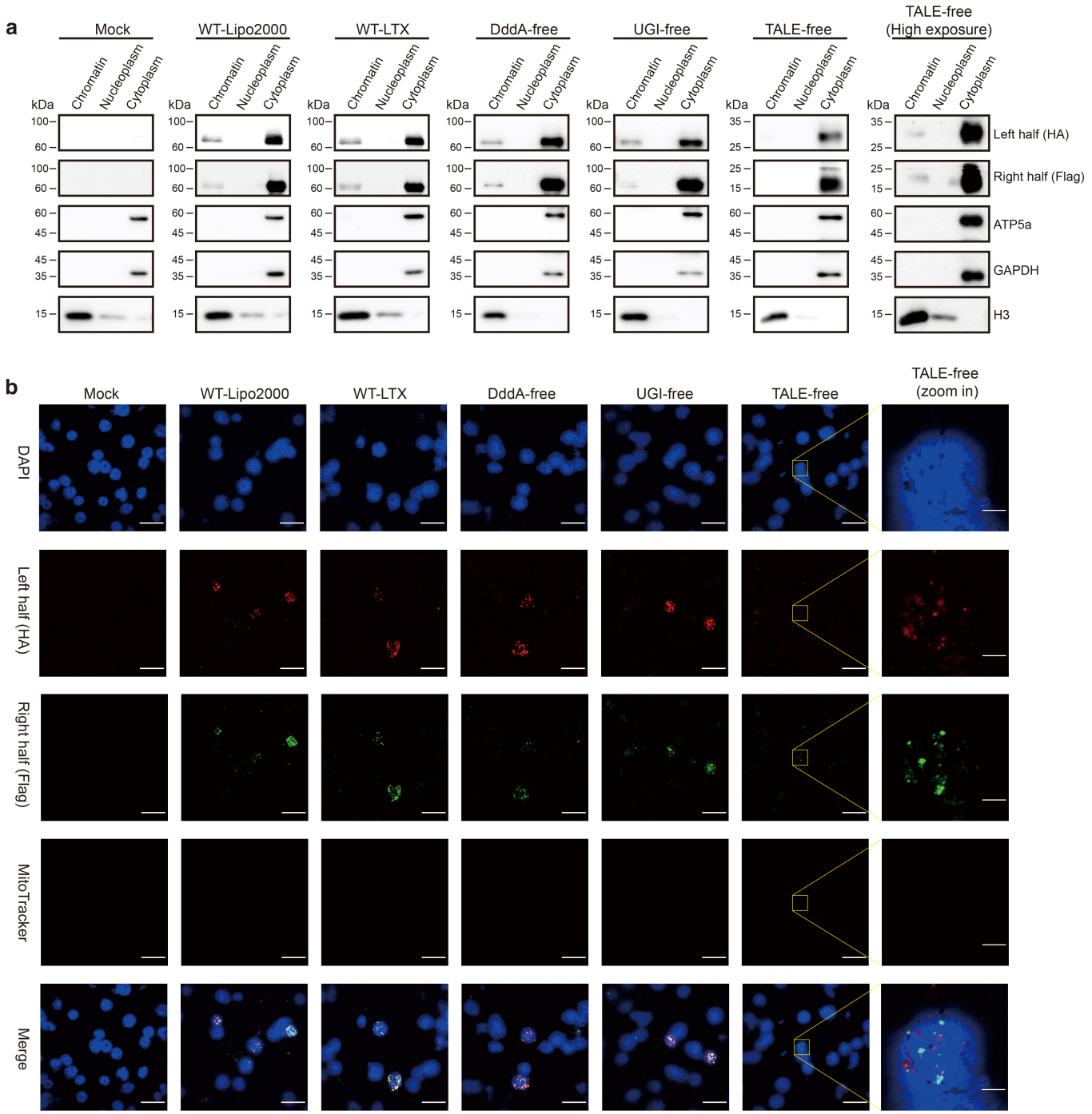


Extended Data Fig. 3 | Editing ratios of nuclear DNA off-target sites identified for the three L1397N-DdCBEs. a–**c**, Targeted deep sequencing results for selected nuclear off-target sites of *ND4*-L1397N (**a**), *ND5.1*-L1397N (**b**) and *ND6*-L1397N (**c**). For each off-target site, the editing ratio for the highest edited cytosine is plotted (blue), and the matched ratio in untreated ctrl sample is plotted in gray.

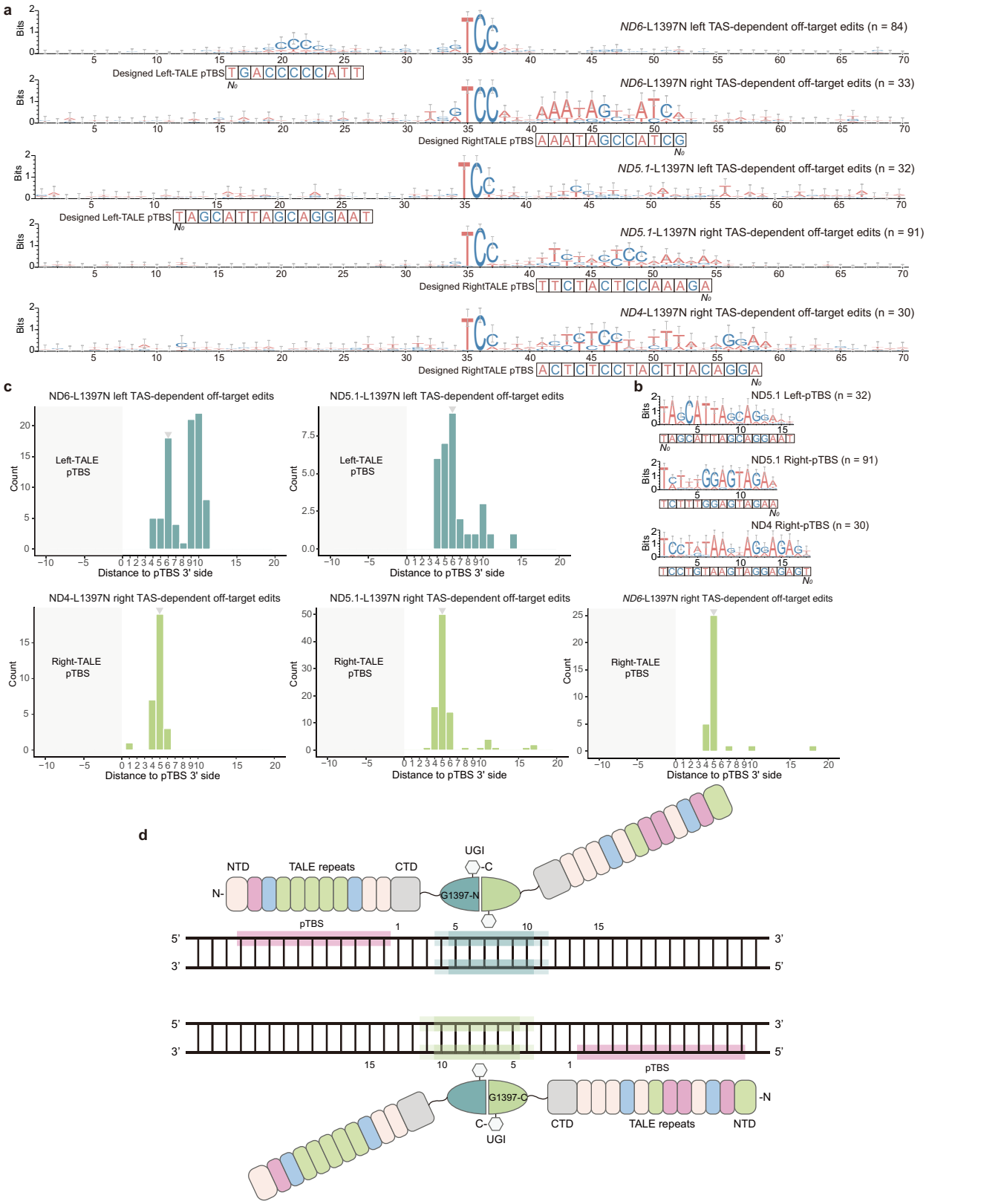


Extended Data Fig. 4 | See next page for caption.

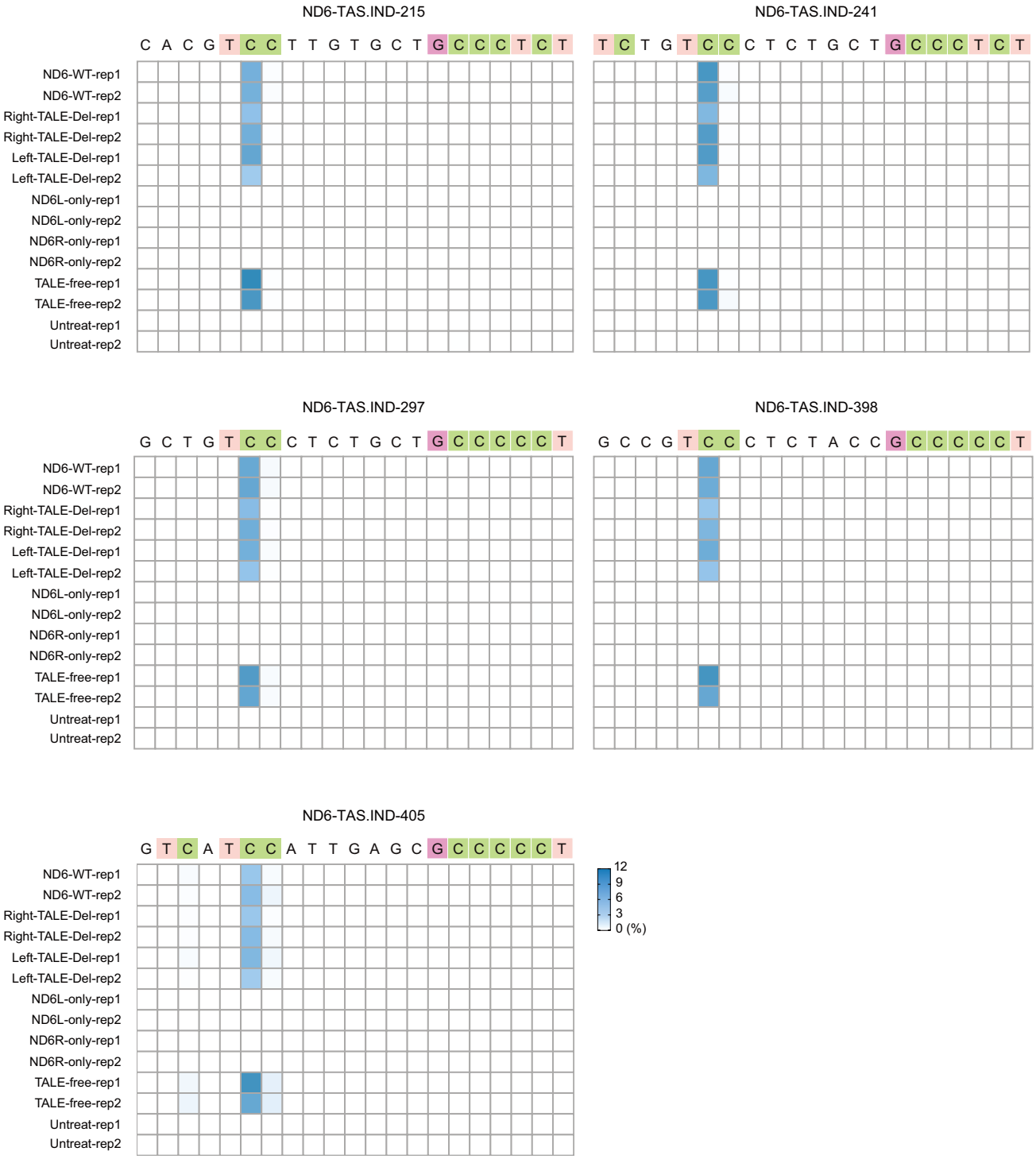
Extended Data Fig. 4 | A real-time IF staining assay using unfixed HeLa nuclei to demonstrate the nuclear localization of DdCBE. a, Fluorescence imaging of DAPI (navy blue), HA-tagged left half (Anti-HA, orange red) and Flag-tagged right half (Anti-Flag, green) in unfixed nuclei of HeLa cells untreated or transfected with Lipofectamine LTX. Possible mitochondrial contamination was tested by MitoTracker (magenta). The images were obtained at a representative Z-axis under the same exposure condition by High Speed Spinning Disk Confocal Microscope (ANDOR). Scale bars, 3 μm. Images are representative of 3 independent biological replicates. **b**, **c**, The projected 2D fluorescence image (**b**) and 3D snapshot (**c**) of a representative nuclei from cells transfected with Lipofectamine 3000. **d**, **e**, The projected 2D fluorescence image (**d**) and 3D snapshot (**e**) of a representative nuclei from cells transfected with Lipofectamine LTX. **f**, Statistic diagram for 3D mean fluorescence intensity per voxel of all scanned nuclei under different treatments. The data in **b**–**f** for each nucleus was obtained from z-stack images collected at 0.4 μm intervals under the same exposure condition by DeltaVision OMX SR (GE). Similar color and scale bars in **a** were used. HeLa cells on 6-well plates were transfected with 2 μg of each monomer using 6 μl Lipofectamine3000; or, cells were transfected with 3.5 μg of each monomer using 21 μl Lipofectamine LTX. "ND6-WT": wild type *ND6*-L1397N; "ND6-(TALE-)": *ND6*-L1397N architectures that deleted the TALE arrays. In **f**, error bars reflect the mean +/− SD; and p-values are calculated by one-side Student's t-test.



Extended Data Fig. 5 | A small portion of DdCBE is localized in the nucleus of transfected HEK293T cells. a, Western blotting results showing the distribution of *ND6*-L1397N (WT) in different subcellular fractions of 2×10^3 HEK293T cells untreated or transfected using Lipofectamine 2000 or LTX; and the distribution of three deletion variants of *ND6*-L1397N in different fractions of cells transfected with LTX. "DddA-free", "UGI-free" and "TALE-free" mean the deletion of DddA, UGI and TALE arrays from the full-length ND6-L1397N respectively. The results show that *ND6*-L1397N is partially localized in the chromatin fraction no matter which transfection reagent was used. The signal of the TALE-free construct in the chromatin fraction is only present when the exposure time is extended. This observation suggests that compared to DddA and UGI, the TALE arrays most strongly affect the nuclear localization. ATP5a (mitochondria), GAPDH (cytosolic) and H3 (chromatin) were chosen as compartment-specific markers, demonstrating the purity of each subcellular fraction. HA (tagged left half) and Flag (tagged right half) were used to indicate the localization of DdCBEs. Molecular weight is given in kDa; images are representative of 2 independent biological replicates; samples are derived from the same batch of experiment and gels were processed in parallel. **b**, Fluorescence imaging of nuclei (DAPI, blue), HA-tagged left half (Anti-HA, red), Flag-tagged right half (Anti-Flag, green) in fixed nuclei isolated from HEK293T cells untreated or transfected with ND6-L1397N (WT) using Lipofectamine 2000, or transfected with *ND6*-L1397N (WT), DddA-free, UGI-free and TALE-free constructs using Lipofectamine LTX. Possible mitochondrial contamination was tested by MitoTracker (magenta). The results show that a small portion of DdCBE is localized in nuclei, regardless of the transfection conditions. TALE arrays more strongly affect the nuclear localization compared with DddA and UGI. Scale bars, 5 μm for zoomed in images of TALE-free; 40 μm for all remaining images. The images were obtained under the same exposure condition and are representative of 2 independent biological replicates.

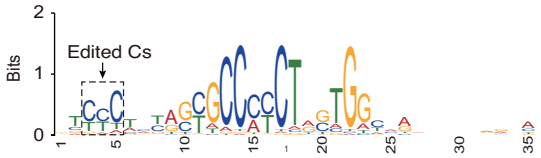


Extended Data Fig. 6 | The editing spectra of DdCBE at TAS-dependent nDNA off-target sites. a, Sequence logos for Cs with highest Detect-seq signal obtained via WebLogo using DNA sequences at TAS-dependent off-target sites of *ND6*-L1397N, *ND5.1*-L1397N and *ND4*-L1397N. **b**, Sequence logos generated from the pTBSs of *ND5.1*-L1397N and *ND4*-L1397N. Bits reflect the level of sequence conservation at a given position. **c**, Aggregate distribution of C·Gs with highest Detect-seq signal across the flanking region of each pTBS for TAS-dependent off-target sites of *ND6*-L1397N, *ND5.1*-L1397N and *ND4*-L1397N. The position of pTBS for left or right TALE proteins is shadowed. **d**, A schematic illustrating the editing spectra of the three L1397N DdCBEs based on the pTBS-edits distribution analysis. Counting the first base pair after the 3' ends of pTBS as position +1. NTD, N-terminal domain; CTD, C-terminal domain.



Extended Data Fig. 7 | The TALE independency of TAS-independent off-target sites validated by targeted deep sequencing. Results of targeted deep sequencing at five representative TAS-independent off-target sites for different *ND6*-L1397N constructs in Fig. 2a.

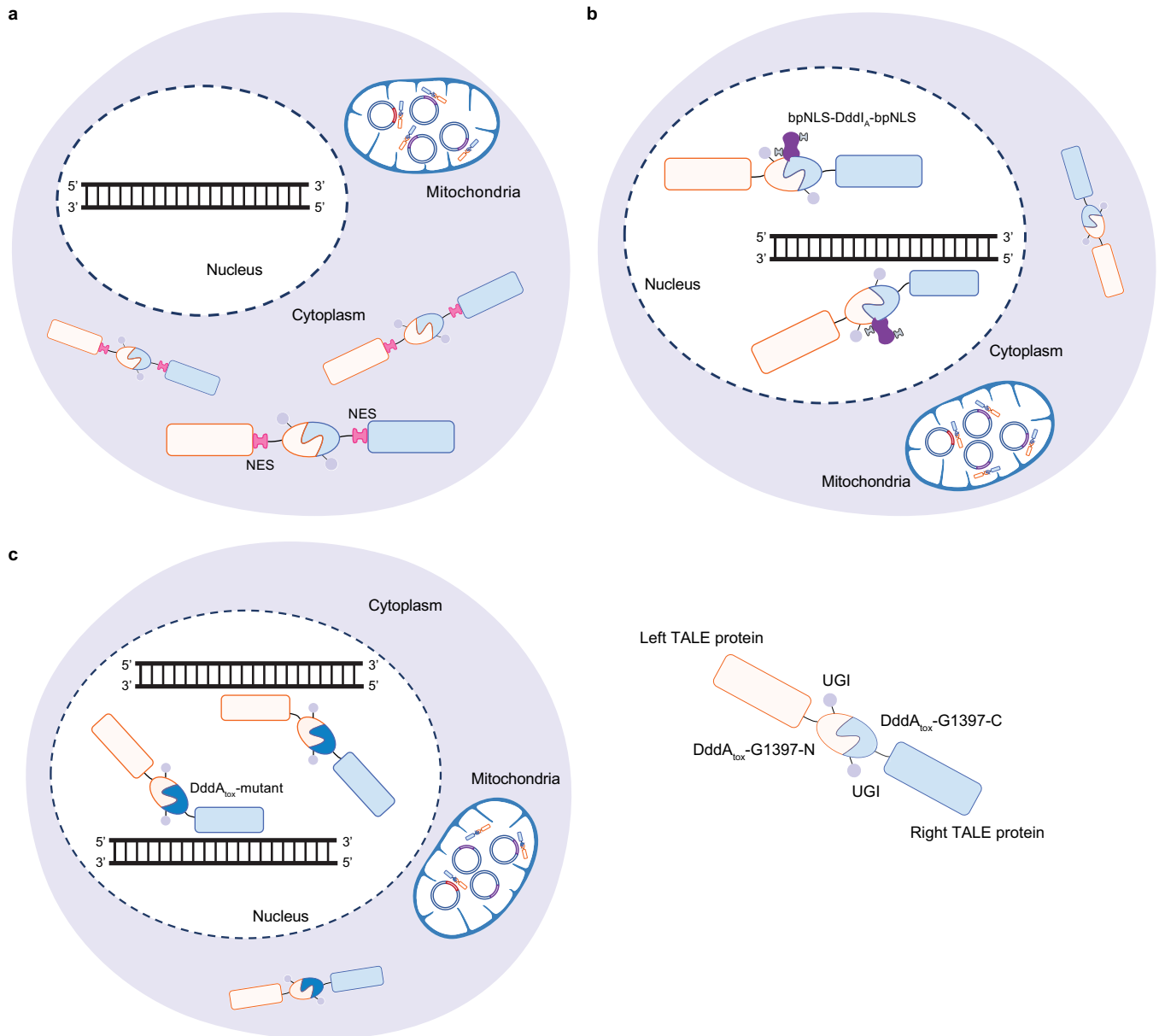
Enriched motif in TAS-independent off-target sites



Motif search result by Tomtom

Name	q-value	E-value	Motif
CTCF	5.34e-11	2.68e-11	
CTCFL	4.77e-08	4.78e-08	
ZIC3	5.15e-01	1.01e+00	
E2F1	5.15e-01	1.05e+00	
SNAI1	5.15e-01	1.51e+00	
KLF8	5.15e-01	1.76e+00	
NR1H3	5.15e-01	1.81e+00	
TFDP1	6.14e-01	2.78e+00	

Extended Data Fig. 8 | Motif search result from sequences of all TAS-independent off-target sites. The results (with a p-value < 0.05) are generated by Tomtom program with JASPAR core motif database.



Extended Data Fig. 9 | Strategies to improve the specificity of DdCBE. **a**, Fusing nuclear export signal (NES) sequences into the DdCBE constructs. The protein level of DdCBE in the nucleus should be decreased, and hence lower the risk of nDNA off-target editing. **b**, Co-expressing DddI$_A$ that fused with nuclear localization signals (NLS). DddI$_A$ is a natural immunity protein of the deaminase DddA; bpNLS-linked DddI$_A$ is supposed to antagonize the deamination activity of DdCBEs mis-localized into the nucleus. bpNLS, bipartite NLS at both the N and C termini. **c**, Mutating the DddA$_{tox}$ in the DdCBE architecture to reduce its intrinsic DNA binding affinity. Ideally, mutated deaminase would not be able to catalyze DNA substrates without the help of simultaneously stable binding of the two TALE repeats.

nature portfolio

Corresponding author(s): Chengqi Yi

Last updated by author(s): Apr 21, 2022

Reporting Summary

Nature Portfolio 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 Portfolio 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	Next-generation sequencing data was collected and demultiplexed by Illumina NovaSeq and Huada MGI platform.
Data analysis	All custom code used for analysis are available at https://github.com/menghaowei/Detect-seq Cutadapt (version 1.18) Bismark (version 0.22.3) BWA MEM (version 0.7.17) samtools (version 1.9) Picard (version 2.0.1) GATK (version 3.8.1) Bowtie2 (version 2.4.2) MACS2 (version 2.1.0) deepTools (version 3.1.3) HiC-Pro (version 3.00) HiCExplorer (3.6) R environment (version 3.6) Bedtools (version 2.27.1) VarScan2 (version 2.4.4) Fiji (version 2.1.0) BD FACSDiva (Version 8.0.1) FlowJo (Version 10.0.7r2)

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All raw sequencing data generated for this paper has been deposited at NCBI GEO and is available under accession number GEO: GSE173859, GSE173689 and GSE176089. The reference genome version used in this study is the human genome 38 (hg38). The Hi-C, DNase-seq, Bisulfite-seq and ChIP-seq data used in this study were downloaded from the GEO or ENCODE database. The GEO accessions are GSE44267, GSM3463661 and GSM3463658. And the ENCODE accessions are ENCFF120XFB, ENCFF993NDR, ENCFF480MMN, ENCFF400CMC, ENCFF449FCR, ENCFF567FDM, ENCFF022DJJ, ENCFF577AJC, ENCFF049YWG, ENCFF273OKN, ENCFF732FSV, ENCFF995CPW, ENCFF267EGW, ENCFF888QBG, ENCFF282XMU, ENCFF268JCB, ENCFF066MYJ, ENCSR022QUM, ENCSR458MAV, ENCSR559QOU, ENCSR699ETV, ENCSR462KQY, ENCSR108ESU, ENCSR224CTR, ENCSR305VIT, ENCSR344YUA, ENCSR113KSF, ENCSR128RMY, ENCSR129YRJ, ENCSR317JGM, and ENCSR403DTW. The detailed information of experiments is available in the 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	For all experiments performed in cell line level, a minimum of 2 (n>=2) biological replicates were performed to confirm reproducibility. In vitro biochemical experiments were performed at least 2 (n>=2) independent times. Our results show that it's sufficient to yield reproducible mean results values. So two biological replicates are sufficient to support conclusions in this paper.
Data exclusions	No data was excluded.
Replication	We performed biological replicates independently at intervals ranging from weeks to months between experiments. All experiments were repeated at least once. All attempts were successful.
Randomization	Not relevant to these experiments.
Blinding	Blinding was not performed as experimental conditions were evident.

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

Antibodies

Antibodies used	rabbit anti-HA (Abcam, ab9110, 1:1000 dilution in TBST for WB; 1:100 or 1:200 dilution in PBS for IF); mouse anti-Flag (Sigma-Aldrich, F1804, 1:2000 dilution in TBST for WB and 1:100 dilution in PBS for ChIP-seq and IF); rabbit anti-CTCF (Abcam, ab128873, 1:2000 dilution in TBST); Goat Anti-Mouse IgG, HRP Conjugated (CWBIO, CW0102, 1:5000 dilution in TBST);

Goat Anti-Rabbit IgG, HRP Conjugated (CWBIO, CW0103, 1:5000 dilution in TBST);
mouse anti-ATP5a (Abcam, ab14748, 1:2000 dilution in TBST);
mouse anti-GAPDH (CWBIO, CW0100, 1:2000 dilution in TBST);
mouse anti-H3 (EASYBIO, BE3015, 1:10000 dilution in TBST) ;
Alexa Fluor 568 Goat anti-Rabbit IgG (Thermo, A-11036, 1:100 dilution in 5% BSA/PBS or 1:500 dilution in 5% FBS/PBS),
Alexa Fluor 488 Goat anti-Mouse IgG (Proteintech, SA00006-1, 1:100 dilution in 5% BSA/PBS)
Alexa Fluor 488 Goat anti-mouse IgG (Thermo, A32723, 1:500 dilution in 5% FBS/PBS)
mouse anti-HA (Abcam, ab1424, 1:2000 or 1:5000 dilution in TBST)
Normal Rabbit IgG (Biodragon, BF01001, 2 µg in 700 µl Co-IP incubation system)
donkey anti-mouse-Alexa 488 (Invitrogen, A21202, 1:500 dilution in PBS).

Validation

rabbit anti-HA: validated by manufacturer by western blotting against cell lysates from 293FT cells transfected with 15kDa HA tagged Vpr (an HIV1 accessory protein) ;
mouse anti-Flag: validated by manufacturer by Immunofluorescence against FLAG tagged myr-PKCz for MDCK canine kidney epithelial cells;
rabbit anti-CTCF: validated by manufacturer by western blotting against cell lysates from HeLa and 293T whole cell lysates;
HRP Conjugated goat Anti-Mouse IgG and goat Anti-Rabbit IgG: Conjugates have been been specifically tested and qualified for Western blot and ELISA assay applications by manufacturer;
mouse anti-ATP5a: validated by manufacturer by western blotting against whole cell lysates from HepG2 cell line and human liver tissue lysate;
mouse anti-GAPDH: validated by manufacturer by western blotting against whole cell lysates from Hela cell line and mouse heart tissue lysate;
mouse anti-H3: validated by manufacturer by western blotting against whole cell lysates from Hela cell line and mouse brain tissue lysate;
mouse anti-HA: validated by manufacturer by western blotting against HEK293T whole cell lysate over-expressing HA-tagged Rab6;
Alexa Fluor conjugated goat Anti-Mouse IgG, goat Anti-Rabbit IgG and donkey Anti-Mouse IgG: Conjugates have been been specifically tested and qualified for Immunofluorescence and Immunocytochemistry assay applications by manufacturer
Normal Rabbit IgG: validated by manufacturer for use as a negative control in parallel with specific primary antibodies in ELISA, FC, Immunoblotting, IF, IHC, IP.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	source: ATCC; cell lines used: HEK293T cells (ATCC CRL-3216); HeLa cells (ATCC CCL-2)
Authentication	All cell lines are from authenticated manufacturers.
Mycoplasma contamination	Cells tested negative for mycoplasma as detailed in the Methods.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

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=GSE173689
Files in database submission	GSM5416444_293T-ISChIP-DdCBE-WT-rep1_bt2_hg38_rmdup_MAPQ20_NotChrM.BinSize10.bigwig GSM5416445_293T-ISChIP-DdCBE-WT-rep1_bt2_hg38_rmdup_MAPQ20_NotChrM.BinSize10.bigwig
Genome browser session (e.g. UCSC)	Reviewers can download the processed in situ ChIP-seq bigwig files from the GEO database.

Methodology

Replicates	Two biological replicates in situ ChIP-seq experiments are performed in HEK293T cell line.
Sequencing depth	The library was purified and selected for 200-1,000 bp fragments for sequencing. The libraries were sequenced by Illumina NovaSeq 6000 sequencer with paired-end 150bp outcomes. The total numbers of in situ ChIP-seq reads are 29.6M and 32.3M, and the unique mapping numbers of reads are 13.6M and 16.4M respectively for rep1 and rep2.
Antibodies	The primary antibody (anti-FLAG, Sigma-Aldrich, F1804), and the secondary antibody (donkey anti-mouse-Alexa 488, A21202).
Peak calling parameters	macs2 callpeak -c Input.BAM -t Flag.InSituChIP.BAM -f BAMPE -g hs

Data quality	In situ ChIP-seq signals are highly reproducible between two biological replicates of ND6-L1397N (Supplementary Fig. 8 a). And there are 20983 enriched peaks with FDR lower than 0.01 and enriched fold change larger than 5.
Software	We used cutadapt (version 1.18) to remove sequencing adapters and mapped clean reads to reference genome hg38 with bowtie2 (version 2.4.2). The additional settings "--no-mixed --no-unal --no-discordant --dovetail --very-sensitive-local -X 2000" were used for fast and sensitive reads alignment. Next, we used Picard (version 2.0.1) to remove PCR duplication and samtools (version 1.9) view command to select alignments with MAPQ over 20. Then we used MACS2 (version 2.1.0) to identify the enriched peaks with default settings. Finally, peaks with q-value smaller than 0.01 and enrichment fold larger than 5 were considered for downstream analysis. And the correlation heatmap plots were generated by deepTools (version 3.1.3) bamCoverage and plotHeatmap programs with "--normalizeUsing RPKM" settings.

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	Subcellular fractions were prepared using Nuclear/Cytosol Fractionation Kit, the supernatants (cytoplasmic extract) were immediately transferred to a clean pre-chilled tube and saved as the cytoplasmic extract fraction. The pellets (containing the nuclei) were washed twice with 600 µl cold PBS,. Then, the nuclei pellets were followed by incubationted in 300 µl PBS containing 10 µg/ml DAPI for 10 min and subjected to FACS.
Instrument	BD Aria SORP
Software	BD FACSDiva software Version 8.0.1; FlowJo (Version 10.0.7r2).
Cell population abundance	Then the nuclei were subjected to FACS sorting step via DAPI signals. The sorted cleaner nuclei were collected in PBS containing 2% FBS. The density of collected nuclei is about 500,000 cells in 1.5ml.
Gating strategy	The nuclei were subjected to FACS sorting step via DAPI signals, the analysis for DAPI is using the Area and Width parameters on the UV 440/40nm channel on the BD Aria SORP.

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