nature methods

ARTICLES

https://doi.org/10.1038/s41592-021-01172-w

Check for updates

Detect-seq reveals out-of-protospacer editing and target-strand editing by cytosine base editors

Zhixin Lei[1,2,8], Haowei Meng[3,8], Zhicong Lv[3,8], Menghao Liu[1,2], Huanan Zhao[4,5], Hao Wu[1,2], Xiaoxue Zhang[1,2], Lulu Liu[3], Yuan Zhuang[3], Kailin Yin[3], Yongchang Yan[3] and Chengqi Yi[1,3,6,7] ✉

Cytosine base editors (CBEs) have the potential to correct human pathogenic point mutations. However, their genome-wide specificity remains poorly understood. Here we report Detect-seq for the evaluation of CBE specificity. It enables sensitive detection of CBE-induced off-target sites at the genome-wide level. Detect-seq leverages chemical labeling and biotin pulldown to trace the editing intermediate deoxyuridine, thereby revealing the editome of CBE. In addition to Cas9-independent and typical Cas9-dependent off-target sites, we discovered edits outside the protospacer sequence (that is, out-of-protospacer) and on the target strand (which pairs with the single-guide RNA). Such unexpected off-target edits are prevalent and can exhibit a high editing ratio, while their occurrences exhibit cell-type dependency and cannot be predicted based on the sgRNA sequence. Moreover, we found out-of-protospacer and target-strand edits nearby the on-target sites tested, challenging the general knowledge that CBEs do not induce proximal off-target mutations. Collectively, our approaches allow unbiased analysis of the CBE editome and provide a widely applicable tool for specificity evaluation of various emerging genome editing tools.

Genome editing tools, especially the CRISPR system, have shown an alluring prospect for therapeutic applications since their introduction[1–4]. Efficient correction of point mutations by base editors, which directly convert DNA bases at targeted loci, provides exciting tools for genetic diseases[2,3,5,6]. Base editors were created by tethering a base modification enzyme[7–9]—for instance, rat APOBEC1 for a cytosine base editor (CBE) or *Escherichia coli* TadA for an adenine base editor—to a catalytically impaired Cas9 nuclease. CBEs use the deaminase to first catalyze conversions from deoxycytidine (dC) to deoxyuridine (dU) and finally result in dC-to-dT transitions restricted within an editing window in the nontarget strand (typically around positions 4–8, counting the protospacer adjacent motif (PAM) as positions 21–23). This process is usually facilitated by uracil-DNA glycosylase inhibitor (UGI) and the Cas9 nickase. Recently optimized CBE tools have enabled base editing in vivo, with high product purity and editing efficiency[5].

Because base editors do not generate double-stranded breaks (DSBs), they are believed to be safer than Cas9 nucleases and thus hold great promise for clinical applications. There are many examples of base editors as potential therapeutics[5,6]. However, their specificity must be thoroughly addressed before base editor tools are ready for clinical use[1,10]. Off-target editing at both DNA and RNA level have been reported for CBEs[11–16]. To identify DNA off-target sites, Kim et al. reported modified Digenome-seq[13], which is based on treatment of extracted genomic DNA with a recombinant base editor lacking UGI and identifies DNA off-target sites without the native chromatin context. Methods relying on clonally derived systems for off-target identification have recently been developed[12,16], but they are low-throughput, time-consuming and technically challenging. Moreover, these methods have reached discordant conclusions: Kim et al. report that a CBE is highly specific and induces only a limited number of Cas9-dependent off-target sites[13], while Zuo et al. and Jin et al. conclude that CBE off-target mutations are random[12,16]. Although it is unclear what results in such discordant observations, a thorough understanding of the off-target effect is a prerequisite for improvement of base editors and ultimately for therapeutic applications. Thus, an unbiased tool is urgently needed to comprehensively evaluate CBE off-target effects at the genome-wide level.

In this study, we develop Detect-seq (dU-detection enabled by C-to-T transition during sequencing) for the genome-wide identification of CBE-induced off-target sites in cellular context. Detect-seq is based on chemical labeling and enrichment of dU, a direct editing product of CBEs, to trace the in vivo editing events in an unbiased manner. Our study expands the current knowledge of off-target effects of CBE and provides a useful method for the specificity assessment of base editors.

Results

A genome-wide method to assess CBE specificity. CBEs catalyze dC-to-dU conversions and finally result in dC-to-dT transitions[8,9]. We thought it might be viable to trace the in vivo editing events and ultimately off-target sites of CBE by capturing the editing intermediate dU (Fig. 1a). Specifically, when genomic DNA is purified from CBE-edited cells, dU generated by CBE in vivo can be recognized by uracil-DNA glycosylase (UDG). Detect-seq uses an in vitro–reconstituted base-excision repair (BER) reaction to achieve specific labeling of dU[17]. During this process normal dTTP and dCTP are replaced by biotin-dUTP and 5-formyl-deoxycytidine triphosphate (5fdCTP) (Extended Data Fig. 1). Biotin-dUTP allows subsequent biotin pulldown of the dU-containing DNA, while multiple 5fdCs incorporate 3′ to the dU sites and this is expected to result in tandem dC-to-dT transitions through a biocompatible chemical reac-

[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]Department of Chemical Biology and Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. [7]Peking University Genome Editing Research Center, Peking University, Beijing, China. [8]These authors contributed equally: Zhixin Lei, Haowei Meng and Zhicong Lv. ✉e-mail: chengqi.yi@pku.edu.cn

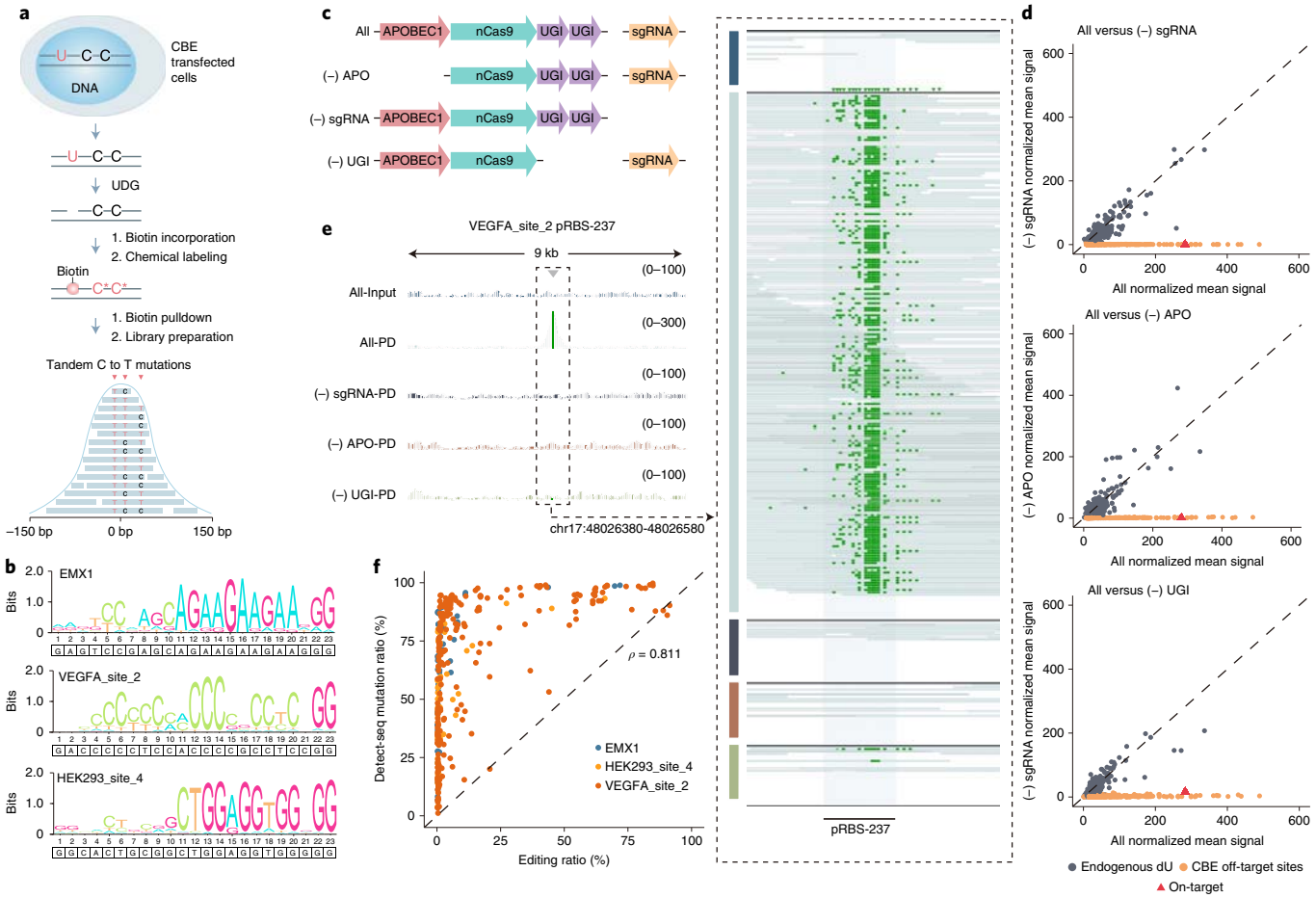

Fig. 1 | Detect-seq assesses the genome-wide specificity of CBE. a, Workflow of Detect-seq. **b**, Sequence logos for EMX1, VEGFA_site_2 and HEK293_site_4 obtained via WebLogo using DNA sequences at the pRBSs. **c**, Architectures of CBE constructs used in the experiments. **d**, Comparisons of Detect-seq signal intensities between architectures in **c**. Endogenous dU (dark gray, $n = 214$) stays on the diagonal, while the on-target site (red) and reproducible off-targets (orange, $n = 511$) respond to the deletion of sgRNA (VEGFA_site_2), APOBEC1 and UGI. **e**, Representative Detect-seq results for a given off-target site of constructs in **c**. The pRBS is shadowed. Green blocks in the views from the Integrative Genomics Viewer indicate C-to-T mutations on the nontarget strand; the green inverted triangles indicate genuine C-to-T edits on the reverse strand according to the results of targeted amplicon sequencing. **f**, Plotting Detect-seq signals against editing ratio obtained by targeted amplicon sequencing. Spearman coefficient is shown. See Supplementary Table 2 for values at individual sites. $n = 2$ biologically independent samples for each group in **d** and **f**. Data shown are generated from the HEK293T cell line unless otherwise specified.

tion[18–20] (Fig. 1a and Extended Data Fig. 1). By finding a consecutive dC-to-dT mutation pattern, we will be able to localize dU with high confidence (Fig. 1a).

To ensure the specificity of our approach, we blocked endogenous 5fdC and also used a damage repair step before the BER labeling reaction to remove endogenous abasic sites (AP), single-strand breaks (SSB) and so on (Extended Data Fig. 1 and Supplementary Fig. 1). Under the optimized conditions, spike-in DNAs containing the dU:dG or dU:dA base pair could be enriched roughly 30–80 fold, while those with a 5fdC:dG, AP:dA or SSB:dA pair were not enriched (Supplementary Fig. 1 and Supplementary Table 1), demonstrating that Detect-seq can efficiently enrich dU-containing DNA and is specific to distinguish dUs from other types of lesion or modification in the genome.

We next applied Detect-seq to off-target evaluation of BE4max in either HEK293T or MCF7 cells for several frequently used sgRNAs—the promiscuous VEGFA_site_2 and human embryonic kidney 293 (HEK293)_site_4, EMX1 and RNF2—with no reported off-target sites[21,22]. As expected, we observed evident peaks with characteristic tandem dC-to-dT mutation patterns at the on-target sites (Extended Data Fig. 2a). These features effectively magnified the signal and could be readily distinguished from genomic background including single nucleotide variations (SNVs) and sequencing errors (Extended Data Fig. 2b,c and Supplementary Fig. 2), hence greatly improving the sensitivity of detection comparing to the whole-genome sequencing (WGS)-based methods. As an important control, such featured signals were entirely absent when the sgRNA was absent or when using a nontarget sgRNA (Extended Data Fig. 2a).

Detect-seq sensitively and unbiasedly profiles CBE editome. We then developed a bioinformatic pipeline to identify Detect-seq patterns throughout the whole genome (Supplementary Fig. 3 and see Methods). We searched for putative sgRNA binding sites (pRBS) within the genomic loci identified by Detect-seq, and identified dozens to hundreds of pRBS-containing loci for EMX1, HEK293_site_4 and VEGFA_site_2 (Fig. 1b and Extended Data Fig. 3a–c; Supplementary Table 3; see Methods). These loci exhibit strong Detect-seq signals, which are highly reproducible among replicates (Extended Data Fig. 3c–f). We then systematically performed Detect-seq for CBE architectures without sgRNA, APOBEC1 and UGI (notated as '(−) sgRNA', '(−) APO' and '(−) UGI', respectively)

(Fig. 1c). We found that Detect-seq signals for these pRBS-containing loci either dropped to background level when sgRNAs or APOBEC1 were omitted, or responded to a varying degree to UGI deletion (Fig. 1d,e and Supplementary Fig. 4), demonstrating that these pRBS-containing loci are typical Cas9-dependent off-target sites. Moreover, using an optimized targeted amplicon sequencing approach with a detection limit of ~0.005%[23,24] (Extended Data Fig. 4a,b), we successfully verified 49/49 EMX1 sites, 51/51 VEGFA_site_2 sites, and 43/43 HEK293_site_4 sites with high, medium, low and no Detect-seq signals, and validated their sgRNA dependency (Fig. 1f, Extended Data Fig. 4c–e, and 13a, Supplementary Tables 1 and 2). Among them, 22 and 116 Cs out of 138 and 458 verified Cs for EMX1 and VEGFA_site_2 respectively showed greater than 5% editing ratio; the most severe off-target mutation showed an editing ratio of greater than 50% and 80% for EMX1 and VEGFA_site_2 respectively, in comparison to a concomitant on-target editing ratio of ~73% and ~84%. Notably, Detect-seq signals could also be used to estimate the in vivo editing levels of CBE (Supplementary Fig. 5).

On the other hand, we were not able to find a reasonable pRBS for many genomic loci identified by Detect-seq. Their Detect-seq signals are usually weak compared to those of the Cas9-dependent off-target sites (Supplementary Fig. 6a). Also, these off-target sites were abundant in the 'All' samples (those with the complete CBE machineries) and the (−)sgRNA samples, but were reduced to the background level in the (−) APO samples (Supplementary Fig. 6b,c). Sequence analysis showed that this type of off-target site presented an obvious TC sequence motif, which matched the preferred sequence context of APOBEC enzymes and disappeared in the (−) APO and control samples (Supplementary Fig. 6d)[25]. Moreover, such off-target mutations seemed to prefer to reside in transcribed regions (Supplementary Fig. 6e,f). These features together suggest that they are the 'random', Cas9-independent off-target sites caused by the overexpression of rAPOBEC1 (refs. [12,16,26]). Taken together, CBE induced both Cas9-dependent and Cas9-independent off-target edits. It is worth mentioning that a reanalysis of genome-wide off-target analysis results[16], which reported only Cas9-independent off-target sites, discovered reproducible Cas9-dependent off-target sites as well (Supplementary Fig. 7), hinting at incomplete data interpretation by the original study.

Comparison of Detect-seq with existing methods. Because several existing methods have reported Cas9-dependent off-target sites for the sgRNAs used in this study, we next compared off-target sites identified by Detect-seq with them. For the RNF2 sgRNA, Detect-seq confirmed its high specificity with no Cas9-dependent off-target sites, consistent with the observation from cell-based GUIDE-seq and DISCOVER-Seq that were developed for detecting Cas9-induced DSBs (Supplementary Fig. 4a)[21,22]. For the VEGFA_site_2 and EMX1 sgRNA, Detect-seq identified most of the off-target sites reported by GUIDE-seq, while it identified around half of them for HEK293_site_4 (Fig. 2a and Supplementary Table 3). In addition, Detect-seq identified many more off-target sites not been reported by GUIDE-seq, especially for the VEGFA_site_2 sgRNA. We then used targeted amplicon sequencing to interrogate off-target sites reported only by Detect-seq, shared by Detect-seq with GUIDE-seq and those not identified by our method. Within the 54 sites that were successfully amplified, we found no evidence of CBE-induced-edits for the 15/17 sites that only detected by GUIDE-seq, while 41 unique off-target sites by Detect-seq were successfully validated (Fig. 2b,c and Supplementary Figs. 8 and 9). Last, all the tested and shared off-target sites were also proved (Fig. 2b and Supplementary Fig. 8).

We next compared Detect-seq results with WGS-based methods. For the shared off-target sites, we found much stronger signals in Detect-seq when compared to WGS and Digenome-seq (Fig. 3a, Extended Data Fig. 2c and Supplementary Fig. 10). Again, Detect-seq identified most of the off-target sites by Digenome-seq, while we discovered a much greater number of unique off-target sites (Fig. 3b and Supplementary Table 3). Analogously, we examined all of the off-target sites detected by Digenome-seq for EMX1 and part of them for HEK293_site_4 through targeted amplicon sequencing. Thirty-five sites were successfully amplified, among which all of the shared off-target sites were proved to be true (Fig. 3c and Supplementary Fig. 11). For sites detected by Digenome-seq only, the targeted sequencing result demonstrated that ten out of 15 presented no genuine editing events, while the remaining five sites exhibited an editing ratio just slightly higher than the background level. Closer examination of Detect-seq data for these sites revealed supporting signals, which, however, did not pass our bioinformatic threshold.

We also compared Detect-seq results with Cas-OFFinder[27], a widely used in silico prediction software for Cas9-induced DSBs. Roughly 30–50% of Detect-seq reported off-target sites were also supported by Cas-OFFinder (Fig. 3d), while many of the rest sites contain either gaps or too many mismatches that were difficult to predict with Cas-OFFinder, even when one or two orders of magnitude more off-target sites were predicted (allowing no more than five mismatches).

To understand the different results with the above methods, we further analyzed potential features of the unique off-target sites reported by other methods. We first performed motif analysis but found generally similar sequence logos for the shared and unique off-target sites (Supplementary Fig. 12a). However, off-target sites only reported by Cas-OFFinder and Digenome-seq lacked active histone marks and open chromatin signals by ATAC-seq (Supplementary Fig. 12b,c), when compared to off-target sites by Detect-seq. This observation indicates that the difference could be due to the lack of consideration for native chromatin state by Cas-OFFinder and Digenome-seq. Altogether, these comparisons demonstrate the specificity and sensitivity of Detect-seq in off-target identification of CBEs.

Prevalent out-of-protospacer editing and target-strand editing. Unexpectedly, we also observed evident Detect-seq signals outside the pRBSs (Fig. 4a). Targeted amplicon sequencing not only proved that they are genuine off-target mutations, but also demonstrated that these edits are dependent on the CBE complex (Supplementary Fig. 13a and Supplementary Table 2). Such signals can be several bases or more than a hundred bases away from the pRBS (Fig. 4a and Supplementary Fig. 13), suggesting that CBE can edit bases far away from the canonical editing window. Out-of-protospacer edits were prevalent as well: nearly half of the typical Cas9-dependent off-target sites identified by Detect-seq were flanked by out-of-protospacer edits (Fig. 4b). One site verified by targeted amplicon sequencing showed an editing ratio of roughly 7.4% (Supplementary Fig. 13c), demonstrating that out-of-protospacer editing can certainly lead to severe biological consequences. In another notable example, a previously believed safe site of CBE turned out to be a bona fide off-target site, whose edited Cs were located upstream of the protospacer (Supplementary Fig. 13d). In fact, this site ranked at the top of off-target sites for Cas9 nuclease[21]. We thus conclude that cytosines outside the protospacer can be edited by CBE.

CBE is supposed to edit only the PAM-containing strand (or non-target strand), but not the sgRNA-pairing strand (or target strand). We also observed evident Detect-seq signals on the target strand (Fig. 4c). While edited Cs on the target strand could be localized to the region paired with the sgRNA, most edited Cs were found outside the region (Extended Data Fig. 5). Target-strand edits were further confirmed by targeted amplicon sequencing; among the verified sites, we observed editing ratios of up to 6.3% (Supplementary Fig. 14a). For this particular site, the ratio of target-strand editing events, which are located out of the protospacer, is comparable with

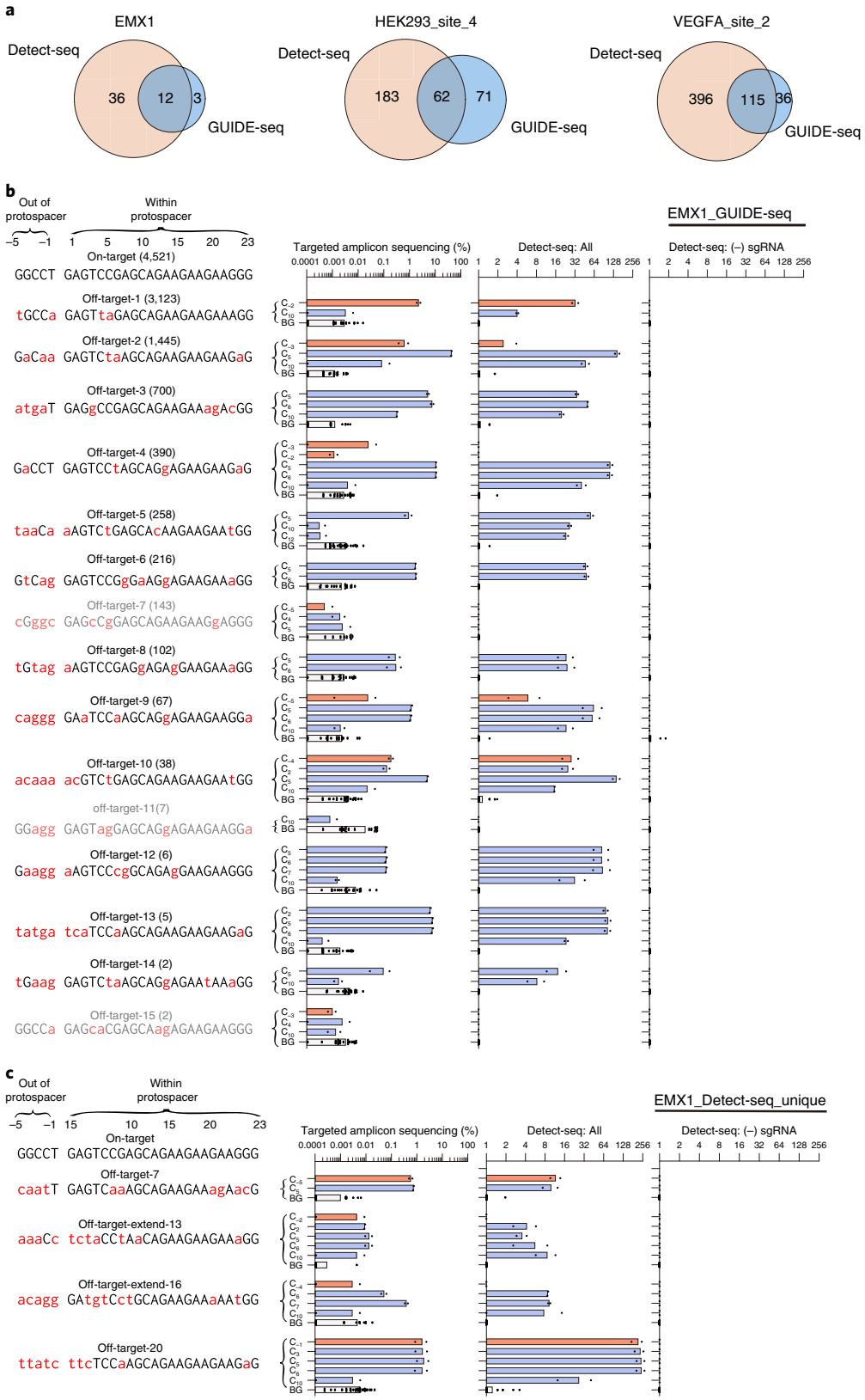

Fig. 2 | Comparisons of Detect-seq results with GUIDE-seq results. a, Venn diagrams of off-target sites by GUIDE-seq and Detect-seq for the three sgRNAs. **b**,**c**, Matched data of Detect-seq and targeted amplicon sequencing for all of the off-target sites of EMX1 by GUIDE-seq (**b**) and unique off-target sites by Detect-seq (**c**). The off-target sites not reported by Detect-seq are marked in gray. In the histograms, cytosines outside the protospacer, cytosines within the protospacer, and background (BG) cytosines are in orange, blue, and gray, respectively. The numbers in parentheses in **b** are GUIDE-seq sequencing read counts.

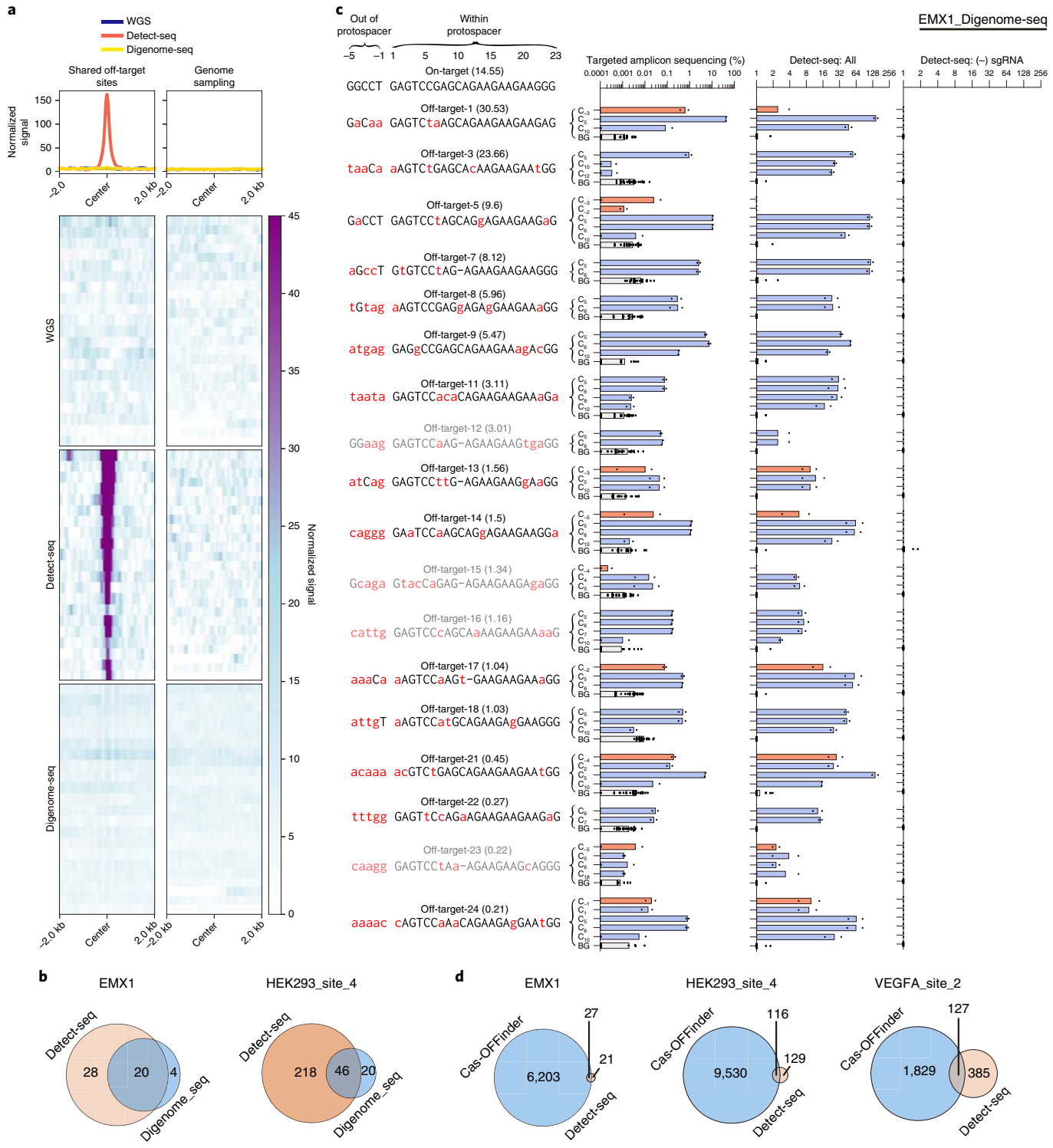

Fig. 3 | Comparisons of Detect-seq with WGS-based methods and computational prediction methods for off-target identification. a, Heatmaps of normalized signals in WGS, Detect-seq and Digenome-seq within a 4 kb window. The left panel shows signals at shared off-target sites by Detect-seq and Digenome-seq, while the right panel shows genome sampling data. **b**, Venn diagrams of off-target sites identified by Digenome-seq and Detect-seq for the two sgRNAs. **c**, Matched results of Detect-seq and targeted amplicon sequencing for the off-target sites of EMX1 by Digenome-seq. All of the off-target sites reported by Digenome-seq are shown except for four sites that were not successfully amplified; the off-target sites not reported by Detect-seq are marked in gray. In the histograms, cytosines outside the protospacer, cytosines within the protospacer, and background (BG) cytosines are in orange, blue, and gray, respectively. The numbers in parentheses are DNA cleavage scores from Digenome-seq. **d**, Venn diagrams of off-target sites identified by Cas-OFFinder (allowing no more than five mismatches) and Detect-seq for the three sgRNAs.

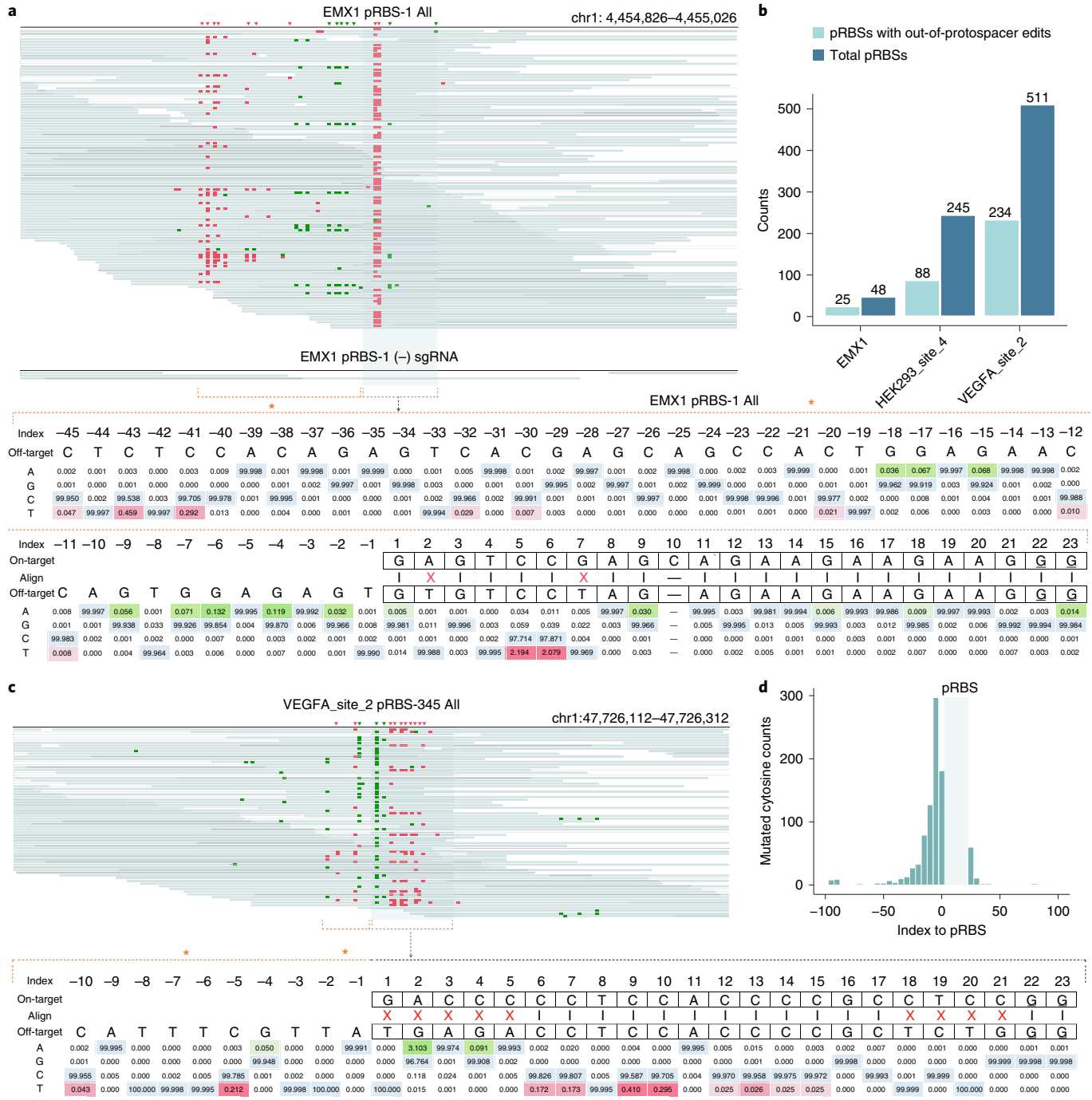

Fig. 4 | Detect-seq discovered prevalent out-of-protospacer edits and target-strand edits. a, A representative example of CBE edits beyond the protospacer. The pRBS is shaded and orange asterisks indicate the region with out-of-protospacer edits. Green blocks indicated target-strand edits; red blocks indicate C-to-T mutations on the nontarget strand; red and green inverted triangles indicate genuine C-to-T edits on the forward and reverse strand, respectively, according to the results of targeted amplicon sequencing. **b**, The frequencies of out-of-protospacer editing events for all aligned pRBSs. **c**, A representative example of CBE edits on the target strand (or sgRNA-pairing strand). Colors and symbols match those in **a**. **d**, The intensity distribution of Detect-seq signals of VEGFA_site_2 at the 5′ and 3′ out-of-protospacer regions. The PAM is counted as positions 21–23.

the highest editing window edits in both HEK293T and MCF7 cells. In total, we identified 11, 20 and 32 loci with target-strand edits for EMX1, VEGFA_site_2 and HEK293_site_4, respectively.

Prompted by the observations that out-of-protospacer and target-strand edits prevalently occur nearby typical Cas9-dependent off-targets, we next interrogated the on-target sites for such edits. Indeed, we were able to find both out-of-protospacer and target-strand edits for all the on-targets tested in this study: the edited Cs can be located a few dozens of bases away from the on-target sites, exhibiting low but evident (up to roughly 0.5%) editing ratio by targeted amplicon sequencing (Supplementary Fig. 15). Even for the RNF2 sgRNA that has no documented off-target sites in literature (Supplementary Fig. 15b), we observed clear out-of-protospacer edits and target-strand edits adjacent to its on-target site. These observations challenge the current knowledge that CBEs typically do not induce proximal off-target edits[5,6,8].

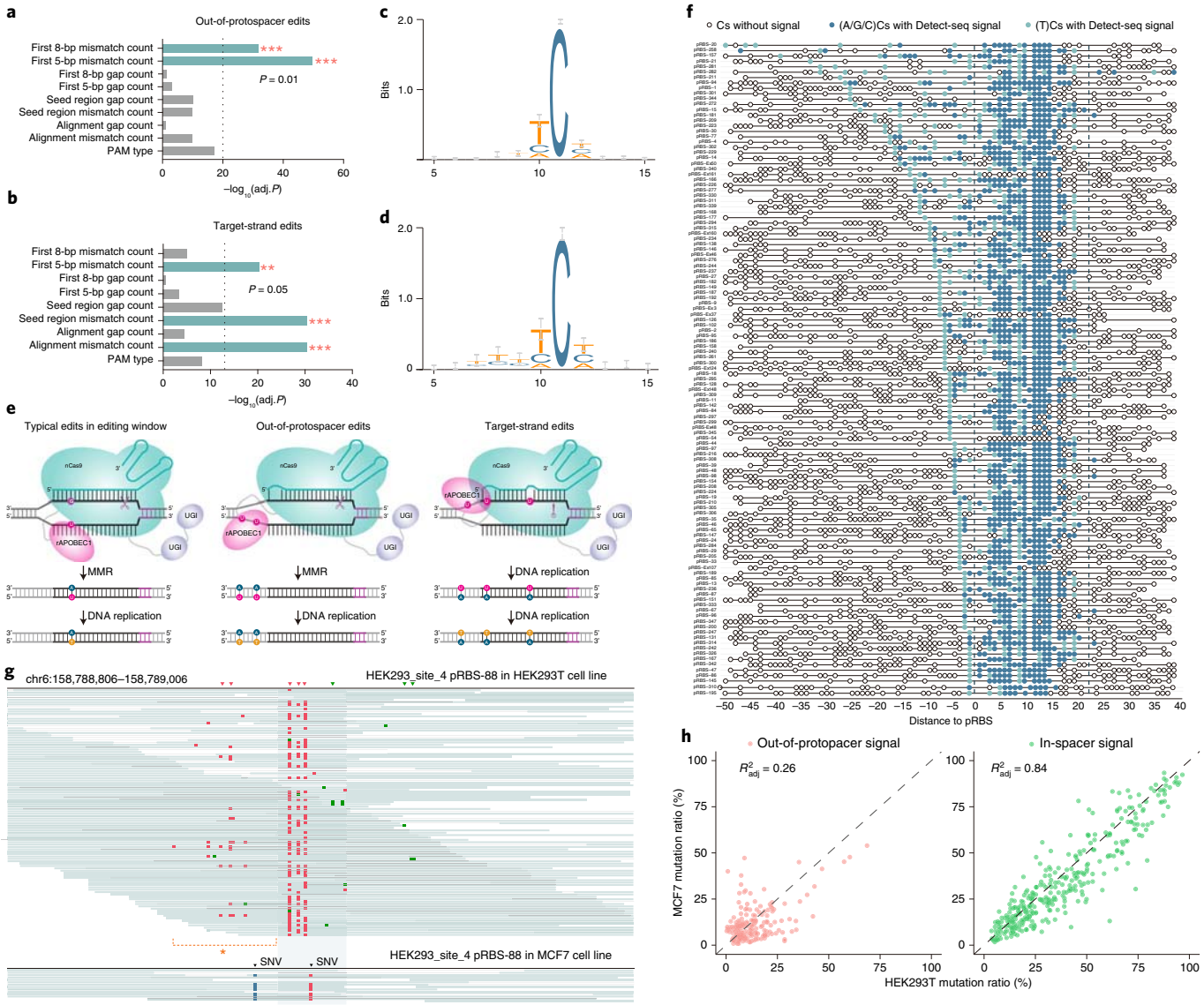

Fig. 5 | The characteristics of out-of-protospacer edits and target-strand edits. a,**b**, Effects of mismatch/gap count and position and PAM type on out-of-protospacer editing (**a**) and target-strand editing (**b**), estimated by ridge linear regression analysis. *P* values were calculated by two-sided Student's *t*-tests. **c**,**d**, The sequence context of out-of-protospacer editing (**c**) or target-strand editing (**d**). The flanking sequences (10 bp on either side, fixing the mutated cytosine in each case at position 11; $n = 1{,}528$ cases in **c** and $n = 335$ cases in **d**, error bars indicate an approximate Bayesian 95% confidence interval) were extracted from the hg38 reference genome to produce a sequence logo using WebLogo. **e**, Proposed models of CBE-induced out-of-protospacer editing and target-strand editing. **f**, Distribution of edited cytosines for different pRBSs with out-of-protospacer edits on the nontarget strands. pRBSs for off-target sites of VEGFA_site_2 sgRNA are indicated by the dashed lines. The PAM is counted as positions 21–23. **g**, An example of divergent CBE editing results at the same genomic site in HEK293T and MCF7 cells. The pRBS is shaded; orange asterisks indicate the region with out-of-protospacer edits. Red blocks in the upper panel indicate C-to-T mutations on the nontarget strand, while blue and red blocks with a triangle above in the lower panel indicate a T-to-C and G-to-T SNV respectively. The red and green inverted triangles in the upper panel indicate genuine C-to-T edits on the forward and reverse strand, respectively, according to the results of targeted amplicon sequencing. **h**, Comparison of Detect-seq signals between HEK293T and MCF7 cell lines in and outside the protospacer region.

Factors influence out-of-protospacer and target-strand edits. To understand the mechanism of out-of-protospacer and target-strand editing, we performed in-depth analysis of the Detect-seq data. We found stronger Detect-seq signals at the PAM distal side than the PAM proximal side for both out-of-protospacer and target-strand edits (Fig. 4d). This observation hinted at a potential link with a certain property of the PAM distal side. We speculated that its DNA secondary structure may play a role. Results of ridge linear regression analysis illustrated that out-of-protospacer edits at the PAM distal side were highly correlated with mismatch numbers in the first 5–8 bp of the pRBSs (Fig. 5a); both counts of mismatches in PAM distal and proximal regions contributed to target-strand edits (Fig. 5b). It was anticipated that these mismatches would lead to imperfect pairing of sgRNA at the off-target DNA loci as well as a destabilized duplex structure at the PAM distal side, with strands unwound as single-strand DNA that could serve as substrates of APOBEC1 (ref. [28]). Indeed, we found a clear TC motif for out-of-protospacer edits and target-strand edits (Fig. 5c,d). In principle, to preserve the edited cytosines on the target strand, the resynthesis of target strand by the BE3-induced mismatch repair

should not occur; thus, CBE may behave like a BE2 in generating target-strand editing (Fig. 5e).

Despite the association with instability of the RNA/DNA hybrid, it is not sufficient to predict out-of-protospacer and target-strand edits. Not all typical Cas9-dependent off-target sites are accompanied with out-of-protospacer or target-strand edits, while the on-target sites, which are supposed to form a stable R-loop structure[29], all possess proximal off-target mutations. Although those edits that occurred were more likely at the 5′ flanking region of pRBS, an evident pattern to ensure which cytosine was edited was lacking for different pRBSs or individual DNA molecules (Fig. 5f, Extended Data Fig. 5b and Supplementary Figs. 16 and 17). Even for a given locus with such edits, only a subset of cytosines embedded in the TC context were edited. In addition, the CBE editome can vary between different cell lines even for the same sgRNA. Not only can typical Cas9-dependent off-target sites be different between HEK293T and MCF7 cells (Fig. 5g and Supplementary Fig. 14), but conserved Cas9-dependent off-target sites can also be surrounded by differential out-of-protospacer edits (Fig. 5h and Supplementary Fig. 14a–d). Take a newly identified Cas9-dependent off-target site as an example, we observed a highest editing ratio of roughly 8.5 and 22.9% within the editing window in HEK293T and MCF7 cells, respectively, whereas a reverse trend (roughly 5.3 and 1.7% efficiency) was found for the highest out-of-protospacer edit (Supplementary Fig. 14c). In addition, target-strand edits can differ by greater than tenfold in the two cell lines (Supplementary Fig. 14d). Altogether, these observations reveal that off-target edits are influenced by various factors and prediction of off-target occurrence remains challenging.

Reassessing improved CBEs. Recent studies have reported CBE variants with improved specificity at the DNA and RNA level[11,15,26,30]. Among them, YE1 (W90Y + R126E) demonstrated the best performance by several independent studies[26,30]; thus we re-evaluated its off-target effects at typical Cas9-dependent, out-of-protospacer and target-strand editing loci identified by Detect-seq. We normalized the expression level of BE4max-YE1 by fluorescence-activated cell sorting (FACS) so as to allow fair comparison with wild-type BE4max (Supplementary Fig. 18). We found that BE4max-YE1 exhibited notably decreased off-target editing at multiple sites tested (Supplementary Fig. 18a), consistent with its reported improved specificity[26,30]. In addition to reduced typical Cas9-dependent off-target effects, BE4max-YE1 showed ameliorated out-of-protospacer and target-strand editing, presumably as a result of its decreased DNA binding affinity. Nevertheless, we also observed off-target loci where its specificity remains to be improved. For instance, at an off-target site discovered only by Detect-seq, BE4max-YE1 presented comparable editing level with wild-type BE4max (roughly 7.0 and 6.2% versus roughly 8.9 and 8.9%, Supplementary Fig. 18b). At another typical Cas9-dependent off-target site, BE4max-YE1 exhibited an editing ratio of roughly 44.1% (Supplementary Fig. 18a). In addition, BE4max-YE1 induced elevated out-of-spacer and target-strand editing level at several tested sites, as well as a roughly twofold increase of indel frequency at the EMX1 on-target site (Supplementary Fig. 18d).

We also compared Cpf1(Cas12a)-BE system with the Cas9-BE(BE4max) system[31,32]. We profiled genome-wide off-target effect of LbCpf1-BE and Cas9-BE, targeting two genomic sites where their editing windows overlap (Extended Data Fig. 6a). Detect-seq revealed hundreds of off-target sites by Cpf1-BE (949 and 240 pRBSs for RUNX1 and DYRK1A, respectively), while much fewer off-target sites (26 for RUNX1 and 31 for DYRK1A) were identified for Cas9-BE (Extended Data Fig. 6b–d and Supplementary Table 3). These data contrast the notion that Cpf1 nuclease is more specific than Cas9; it is possible that the base editor system may differ from its corresponding nuclease system, which has been documented by literature[13,33]. None of the Cas9-base editor induced off-target sites overlaps with those of Cpf1-BE (Extended Data Fig. 6e); this is expected given their different seed regions and PAM sequences (Extended Data Fig. 6f,g). We also validated 23 sites and proved that off-target sites of Cpf1-BE and Cas9-BE are orthogonal (Supplementary Fig. 19). Collectively, Detect-seq enables evaluation of existing tools and future development of more specific base editors.

Discussion

CBEs are one of the most promising genome editing tools to correct human pathogenic mutations[5,6], but their specificity must be carefully examined before therapeutic applications[1,10]. Previous efforts examining CBE-induced DNA off-target effects have reached inconsistent conclusions, confounding the evaluation and development of more specific CBEs. These discrepancies are likely caused by limitations of the methodology and analysis tools. Moreover, methods based on single clones or embryos can be influenced by clone-to-clone variation, which may range in mutation frequencies from below that of control cells to orders of magnitude higher[14]. More recently, a rapid and cost-effective method for assessing Cas9-independent off-target editing has been reported, but it evaluates the capability instead of profiling genuine off-target sites[26]. In this study, we present a sensitive and unbiased method to characterize the CBE editome inside of cells. We show that CBE induces not only prevalent Cas9-dependent and Cas9-independent off-target sites, but also out-of-protospacer edits and target-strand edits. Such an improvement in understanding of the CBE editome will enable the identification of high-specificity CBE variants in the future.

The unexpected discovery of out-of-protospacer edits and target-strand edits expands the current understanding of CBE off-target effects. They are prevalent and can exhibit a high editing ratio, but are influenced by various biological contexts. In addition, the typical Cas9-dependent off-target sites of a CBE differ from the sites induced by Cas9 nuclease alone and cannot be reliably predicted from the sgRNA sequence[13,21,27]. Therefore, we recommend assessing systemically genome-wide specificity for therapeutic applications where base editors have proved to be effective. This is further exemplified by the unanticipated discovery that Cpf1(Cas12a)-BE was able to induce more off-target edits than Cas9-BE, even though Cas12a is more accurate than Cas9 nuclease[34,35]. We speculate that the higher binding sequence tolerance of Cpf1-BE compared to Cas9-BE or the difference between the nuclease and base editor system might contribute to the unexpected observation.

Detect-seq offers a robust platform for off-target detection. It captures dU to profile the editome of CBE. We believe its applicability goes beyond assessment of CBE. In fact, the recently developed ACBE[36–39], GBE[40,41] and mitochondrial base editor DdCBE[42] all generate dU as an editing intermediate. Hence, Detect-seq is anticipated to have wide applications in specificity evaluation and tool development in the field of genome editing. Additionally, it is in principle applicable to various biological contexts, including postmitotic cells, patient-derived primary cells and animal disease models, where genome editing tools have been used.

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/s41592-021-01172-w.

Received: 6 August 2020; Accepted: 3 May 2021;
Published online: 7 June 2021

References

1. Doudna, J. A. The promise and challenge of therapeutic genome editing. *Nature* **578**, 229–236 (2020).
2. Lee, J. et al. Recent advances in genome editing of stem cells for drug discovery and therapeutic application. *Pharmacol. Ther.* **209**, 107501 (2020).
3. Wang, D., Zhang, F. & Gao, G. CRISPR-based therapeutic genome editing: strategies and in vivo delivery by AAV vectors. *Cell* **181**, 136–150 (2020).
4. Komor, A. C., Badran, A. H. & Liu, D. R. CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell* **169**, 559 (2017).
5. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).
6. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
7. Gaudelli, N. M. et al. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
8. 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).
9. Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, aaf8729 (2016).
10. Dunbar, C. E. et al. Gene therapy comes of age. *Science* **359**, eaan4672 (2018).
11. Grunewald, J. et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).
12. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
13. Kim, D. et al. Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. *Nat. Biotechnol.* **35**, 475–480 (2017).
14. McGrath, E. et al. Targeting specificity of APOBEC-based cytosine base editor in human iPSCs determined by whole genome sequencing. *Nat. Commun.* **10**, 5353 (2019).
15. Zhou, C. et al. Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature* **571**, 275–278 (2019).
16. Zuo, E. et al. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289–292 (2019).
17. Shu, X. et al. Genome-wide mapping reveals that deoxyuridine is enriched in the human centromeric DNA. *Nat. Chem. Biol.* **14**, 680–687 (2018).
18. Xia, B. et al. Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **12**, 1047–1050 (2015).
19. 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).
20. Zeng, H. et al. Bisulfite-free, nanoscale analysis of 5-hydroxymethylcytosine at single base resolution. *J. Am. Chem. Soc.* **140**, 13190–13194 (2018).
21. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
22. Wienert, B. et al. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science* **364**, 286–289 (2019).
23. Hong, J. & Gresham, D. Incorporation of unique molecular identifiers in TruSeq adapters improves the accuracy of quantitative sequencing. *Biotechniques* **63**, 221–226 (2017).
24. Salk, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.* **19**, 269–285 (2018).
25. Saraconi, G., Severi, F., Sala, C., Mattiuz, G. & Conticello, S. G. The RNA editing enzyme APOBEC1 induces somatic mutations and a compatible mutational signature is present in esophageal adenocarcinomas. *Genome Biol.* **15**, 417 (2014).
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. Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
28. Huai, C. et al. Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nat. Commun.* **8**, 1375 (2017).
29. Jiang, F. et al. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**, 867–871 (2016).
30. Zuo, E. et al. A rationally engineered cytosine base editor retains high on-target activity while reducing both DNA and RNA off-target effects. *Nat. Methods* **17**, 600–604 (2020).
31. Li, X. et al. Base editing with a Cpf1-cytidine deaminase fusion. *Nat. Biotechnol.* **36**, 324–327 (2018).
32. Wang, X. et al. Cas12a base editors induce efficient and specific editing with low DNA damage response. *Cell Rep.* **31**, 107723 (2020).
33. Kim, D., Lim, K., Kim, D. E. & Kim, J. S. Genome-wide specificity of dCpf1 cytidine base editors. *Nat. Commun.* **11**, 4072 (2020).
34. Kim, D. et al. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34**, 863–868 (2016).
35. Kleinstiver, B. P. et al. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **34**, 869–874 (2016).
36. Sakata, R. C. et al. Base editors for simultaneous introduction of C-to-T and A-to-G mutations. *Nat. Biotechnol.* **38**, 865–869 (2020).
37. Zhang, X. et al. Dual base editor catalyzes both cytosine and adenine base conversions in human cells. *Nat. Biotechnol.* **38**, 856–860 (2020).
38. Grunewald, J. et al. A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing. *Nat. Biotechnol.* **38**, 861–864 (2020).
39. Li, C. et al. Targeted, random mutagenesis of plant genes with dual cytosine and adenine base editors. *Nat. Biotechnol.* **38**, 875–882 (2020).
40. Kurt, I. C. et al. CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nat. Biotech.* **39**, 41–46 (2020).
41. Zhao, D. et al. Glycosylase base editors enable C-to-A and C-to-G base changes. *Nat. Biotech.* **39**, 35–40 (2020).
42. Mok, B. Y. et al. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* **583**, 631–637 (2020).

Methods

Cell culture. HEK293T (ATCC, CRL-11268) and MCF7 (ATCC, HTB-22) cells were separately maintained in DMEM (Corning, 10-013-CVR) and MEM (Corning, 10-009-CVR) supplemented with 10% FBS (Gibco) and 1% penicillin/streptomycin (Gibco, 10378016) at 37 °C under 5% CO_2. The subculture of cells was performed every 2–3 d and only passages 3–10 were used for experiments. All the cells were routinely tested for Mycoplasma contamination with TransDetect PCR Mycoplasma Detection Kit (Transgen Biotech, FM311-01).

Preparation of spike-in model sequences. Single-strand oligonucleotides were synthesized on an Expedite 8909 DNA synthesizer using standard reagents (Glen Research, Inc.). The control GC model sequence contains only canonical bases, and it was generated by primer extension using EASYTaq DNA Polymerase (Transgen Biotech, AP111) with dATP/dGTP/dCTP/dTTP. To obtain model sequences containing a dU:dA, dU:dG or 5fdC:dG pair, a dU- or 5fdC-containing single-strand DNA and a partially overlapped complementary strand were annealed and extended with EASYTaq DNA Polymerase, followed by an exonuclease I digestion and a purification with 1.8× Agencourt AMPure XP beads (Beckman Coulter). The sequence with a single abasic site was generated from a double-strand DNA sequence containing one dU:dA pair by removal of uracil base with UDG (NEB, M0280) at 37 °C for 1 h followed with a purification by 1.8× Agencourt AMPure XP beads; the sequence with a single SSB site was generated from a double-strand DNA sequence containing one abasic site:dA pair by the cleavage of abasic site with endonuclease IV (Endo IV, NEB, M0304) at 37 °C for 2 h followed with a purification by 1.8× Agencourt AMPure XP beads. All sequences used for above experiments were purchased from Invitrogen (Thermo Scientific). All spike-in sequences (Supplementary Table 1) were purified by 8% native PAGE and finally stocked in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) at −80 °C.

Plasmid cloning. The BE4-deletion variants were constructed with GIBSON assembly mix (NEB) based on the BE4max plasmid, which was a gift from the laboratory of Wensheng Wei. Control experiments were performed using a mCherry-only plasmid that did not contain any base editor domains. sgRNAs were cloned into an expression vector under the control of a U6 promoter using the Golden Gate method. The YE1-BE4max mutants were achieved via a highly efficient point mutation strategy from the BE4max plasmid using TransStart FastPfu DNA Polymerase (Transgen Biotech, AP221-01). To allow fair comparison of dCpf1- and Cas9-based BE systems, the dLbCpf1-BE construct is obtained by replacing the APOBEC3A in the BEACON2 plasmid (a gift from the laboratory of Jia Chen) with rAPOBEC1 from the BE4max plasmid using Gibson assembly.

Transfections. For transfections, 6.4×10^6 HEK293T cells or 2.5×10^6 MCF7 cells were seeded into six-well culture plates (Corning) for 16 h growth. Adherent cells were transfected with 4 μg of base editor and 2,720 ng of sgRNA plasmids per well using lipofectamine LTX following the manufacturer's protocol. 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 TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) at −80 °C.

Detect-seq. Extracted genomic DNA was fragmented into roughly 300 bp through the Covaris Focused-ultrasonicator Instrument (ME220). Roughly 5 μg of DNA fragments and 10 pg of spike-in model sequences were subjected to end repair with NEBNext End Repair Module (NEB, E6050); *E. coli* ligase (NEB, M0205) was also added during this step to remove nicks in the DNA. Then hydroxylamine protection of endogenous 5fdC was performed in 100 mM MES buffer (pH 5.0), 10 mM *O*-ethylhydroxylamine ($EtONH_2$, Aldrich, 274992) at 37 °C for 6 h. dA was added to the 3′ end of DNA by NEBNext dA-Tailing Module (NEB, E6053).

DNA damages that may have interfered the subsequent labeling step were repaired in a mixture of 2 μl of Endo IV (NEB, M0304), 1 μl of *Bst* full-length polymerase (NEB, M0328), 2 μl Taq ligase (NEB, M0208), 1 μl NAD^+ (NEB, B9007), 1 μl of dNTP (2.5 mM each) in NEBuffer 3 for 1 h at 37 °C and 1 h at 45 °C. Note that during this damage repair step, potential signal noise from endogenous abasic sites, single-stranded breaks, nicks and so on are removed. DNA was purified and subjected to in vitro BER labeling reaction containing a mixture of 1 μl of UDG (NEB, M0280), 1.5 μl of Endo IV, 0.8 μl of *Bst* full-length polymerase, 1.7 μl of Taq ligase, 1 μl of NAD^+, 200 nM biotin-dUTP (Trilink, N-5001-050), 800 nM 5fdCTP (Trilink, N-2064-1), 200 nM dATP and 200 nM dGTP in NEBuffer 3 for 40 min at 37 °C. Then DNA was incubated with 75 mM of malononitrile in 10 mM Tris-HCl (pH 7.0) at 37 °C for 20 h in a thermomixer (Eppendorf, 850 r.p.m.).

Labeled fragments were enriched by streptavidin C1 beads (Invitrogen) following the manufacturer's instructions. A Y adapter was ligated (NEBNext Quick Ligation Module, E6056) to double-stranded DNA on streptavidin C1 beads and free adapters were removed by washing three times with 1× B&W buffer (5 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween-20), followed by treatment of NaOH. The DNA was eluted from C1 beads using deionized water after heating at 95 °C for 3 min. Eluted DNA was finally subjected to PCR amplification. Sequencing was performed by Illumina Hiseq X Ten and MGISEQ-2000.

For preliminarily evaluating the efficiency and specificity of Detect-seq, we first compared samples before and after pulldown by using quantitative PCR and Sanger sequencing on spike-in molecules. Specifically, the level of enrichment after biotin pulldown for spike-in sequences (containing dU:dA, dU:dG or other base pairs with DNA modifications) were calculated using the $2^{-\Delta\Delta Ct}$ method normalized by the control GC model sequence. The PCR products of spike-ins and on-target sites were subjected to Sanger sequencing to assess the efficiency of C-to-T conversions and pulldown.

FACS. Cells were washed with 1× PBS (Corning) and treated with 0.25% Trypsin-EDTA (Gibco) solution. Cells were then diluted to a concentration of roughly 1×10^7 cells per ml with Opti-MEM and passed through a 35-μm cell strainer cap (Corning). After gating for the singlet cell population, filtered cells were carried out on a FACS Aria III (BD Biosciences). Cells transfected were sorted for $mCherry^+$ signal and the approximately top 20–50% of cells with the highest signal were collected into prechilled Opti-MEM solution. All the samples were sorted with an equal mean fluorescence intensity of mCherry signal.

Targeted amplicon sequencing. Regions flanking the targeting sites were selected for the design of primers, whose overhangs contained the paired Illumina adapter sequences. In addition, a 10-nt barcode was also added into each primer pair (Supplementary Table 1) to lower the detection limit from 10^{-3} to 10^{-5}. The first round of PCR amplification was performed with NEBNext Q5 Hot Start HiFi PCR Master Mix (NEB, M0543L) using roughly 10–100 ng of genomic DNA as an input template. After about ten cycles of amplification, the PCR products were purified with 1× Agencourt AMPure XP beads and eluted with ddH_2O. Purified DNA samples were then subjected to the second round of amplification for roughly 15 cycles and assigned with different indexes followed by a purification with 0.8× AMPure XP beads. The libraries were quantified with Qubit 2.0 Fluorometer (Invitrogen) and pooled for high-throughput sequencing by Illumina HiSeq X Ten (Genetron Health).

Detect-seq mapping. Illumina sequencing adapters in Detect-seq raw FASTQ files were removed by cutadapt software (v.1.18). Working command and key parameters were as follows: cutadapt --times 1 -e 0.1 -O 3 --quality-cutoff 25 -m 50. After the adapter removal, FASTQ files were mapped to the reference genome (hg38) with a converted sequence reads aligner Bismark (v.0.22.3) with default settings. Then the unmapped and reads with mapping quality lower than 20 were remapped by BWA MEM (v.0.7.17) with default parameters. The Bismark-generated BAM and the BWA-generated BAM files were merged, and the merged BAM files were sorted by reference coordinate with samtools sort command (v.1.9). Duplications were removed from the sorted BAM files by Picard MarkDuplicates (v.2.0.1). Considering CBE usually generates indel byproducts, BAM files were processed by GATK RealignerTargetCreator and IndelRealigner (v.3.8.1) with default settings, and the known single nucleotide polymorphism sites used in this step were downloaded from the GTAK resource bundle (v.NCBI dbSNP 138).

Detect-seq tools and code availability. After the Detect-seq mapping steps, several computation and analysis steps should be performed to obtain the final Detect-seq off-target sites. To make the analysis pipeline easy to implement, we have deposited Detect-seq tools containing several useful Python scripts on GitHub (https://github.com/menghaowei/Detect-seq). Detect-seq tools can help to perform Detect-seq analysis including, but not limited to, tandem C-to-T signal finding, enrichment test, off-target site identification, sgRNA alignment and results visualization.

Identification of regions with Detect-seq signals. To obtain Detect-seq regions with tandem C-to-T feature among the whole genome, first we generated mpileup files from BAM files by samtools mpileup command (v.1.9) with the key parameters -q 20 -Q 20. Then mpileup files were processed to .bmat and .pmat files by Detect-seq tools parse-mpileup and bmat2pmat commands with default settings. We next searched the tandem C-to-T pattern in the whole genome by pmat-merge command and obtained merged .pmat files (so-called .mpmat files). The .mpmat files were filtered with mpmat-select command with settings -m 3 -c 6 -r 0.01 --RegionPassNum 1 --RegionToleranceNum 3. After the .mpmat filtering step, we obtained the preliminary regions with Detect-seq signals. All scripts used in this step were collected into the Detect-seq tools.

Mutation reads, nonmutation reads and count normalization. To normalize the sequencing depth, we calculated the normalized count of Detect-seq signals. Sequencing reads with no fewer than one tandem C-to-T mutations were deemed to be Detect-seq mutation reads, while sequencing reads without a C-to-T mutation were defined as nonmutation reads. The total read count was a summation of mutation read count and nonmutation read count. We normalized read count using the following formula:

$$\text{NormalizedCount} = \frac{\text{Region Raw Count}}{\text{Total Raw Count}/10^6} \times 100$$

We calculated normalized mutation and normalized total read counts, respectively, which were used for the subsequent analysis.

The Poisson test for Detect-seq regions. Referring to a well-known peak calling algorithm MACS, which assumes that the sequencing reads obey the Poisson distribution, we coded a find-significant-mpmat script for Detect-seq enrichment analysis and statistical testing. During this analysis step, each preliminary Detect-seq signal region was calculated for the normalized mutation Detect-seq count for the control sample and treatment sample, respectively. Then a Poisson one-side test was performed and the parameter lambda in this test was set as the normalized Detect-seq mutation read count in the control sample. After the statistical test, the *P* value was adjusted using the Benjamini and Hochberg method to control the false discovery rate (FDR).

Identification of endogenous dU. We searched endogenous dU by comparing the Detect-seq signals between mCherry samples and All-Input samples. We used the Detect-seq tools find-significant-mpmat script to test the enrichment of those potential endogenous dU regions. The region that complied with the following criteria was considered to be an endogenous dU: FDR < 0.05, fold change of normalized mutation reads count in the mCherry sample to normalized mutation read count in the corresponding All-Input should be larger than 1.5, the normalized mutation read count in the mCherry sample to be no less than five and the mutation read counts in the All-Input sample to be no greater than three.

Alignment for the putative sgRNA/crispr RNA binding sites. To find a putative binding site for sgRNA/crRNA (pRBS), we extracted sequences from the reference genome hg38 and aligned them with the on-target sequence by a modified semiglobal alignment algorithm. First, we searched the PAM sequence on both strands of the extracted sequences. For Cas9-BE, we searched all NRG (R stands for A or G) on the extracted sequences and set those motifs as candidate PAMs. While for the Cpf1-BE, we searched TTTV (V stands for A, C or G) and set those motifs as Cpf1 candidate PAMs. Then we extracted 30-nt sequences from the 5′ or 3′ direction related to the candidate PAMs for Cas9-BE or Cpf1-BE, respectively. Next, we ran a standard semiglobal alignment between those candidates and the on-target sequence without PAM. Meanwhile, a directly pairwise semiglobal alignment between extracted sequences and the on-target sequence was performed. The alignment with highest score was reported as the putative sgRNA/crRNA binding stie. The semiglobal alignment parameters were set as match +5, mismatch −4, gap open −24 and gap extension −8.

Identification of the Cas9-dependent off-target sites. We identified Cas-dependent off-target sites by comparing the Detect-seq signals between mCherry samples and All-PD samples. First, we used Detect-seq tools find-significant-mpmat script to test the enrichment of preliminary Detect-seq signal regions. The region that complied with the following criteria was considered to be a candidate Cas-dependent off-target site: FDR < 0.05; fold change of normalized mutation read count in the All-PD sample to normalized mutation read count in the mCherry sample to be greater than two; the mutation read count in the mCherry sample to be no larger than one and the mutation read count in the All-PD sample to be no fewer than five. Second, the candidate Cas-dependent off-target sites were aligned with the on-target sequence by a modified semiglobal alignment algorithm that was described above. Finally, the best alignment was reported as pRBS.

Identifying the Cas9-independent off-target sites. To make a fair comparison among the different data sets, we first downsampled all data sets to the same sequencing depth. Then we searched tandem C-to-T regions among the whole genome in All-PD samples, (−) sgRNA samples and (−) APO samples with the pmat-merge script. After this searching step, we filtered the tandem C-to-T regions by mpmat-select command with settings as -m 2 -c 4 -r 0.01 --RegionPassNum 2 --RegionToleranceNum 0. Next, we aligned the on-target sequence with each tandem C-to-T region in All samples, (−) sgRNA samples and (−) APO samples. The tandem C-to-T regions with an alignment score higher than eight were excluded for the downstream analysis. Finally, we removed the tandem C-to-T regions from All samples, (−) sgRNA samples and (−) APO samples, if they contained any mutation signal in mCherry samples. The remaining regions in each sample were considered to be Cas-independent off-target sites.

Identifying out-of-protospacer edits and target-strand edits. First we considered all Cs on both genomic strands for each pRBS within a 200-bp interval as potential candidates. Then we calculated a *P* value for each candidate C by comparing its mutation read count with the whole-genome mutation background by a one-side binomial test. The *P* values were adjusted with the Benjamini–Hochberg method to control the FDR. Finally, the cytosine that complied with the following criteria was considered to be a real edited C: adjusted $P < 0.01$, mutation read count no less than five and mutation ratio larger than 0.005%.

Effect factor analysis with ridge regression. To find the effective factors in out-of-protospacer edits and target-strand edits, we performed a ridge regression analysis by ridge package (v.2.4) in the R environment (v.3.6). All identified pRBSs in HEK293T cell line were involved in this analysis. For out-of-protospacer analysis, the dependent variable was set as a zero or one binary value, referring to whether or not there were out-of-protospacer edits for each specific pRBS. Then we selected sgRNA alignment mismatch count, gap count at PAM distal side, sgRNA alignment seed region mismatch count, seed region gap count, sgRNA alignment total mismatch count, sgRNA alignment total gap count and sgRNA PAM type as the independent factors. Next, we fitted a linear ridge regression model between the dependent variable and the selected independent variables. Finally, the factor with a *P* value lower than 0.05 was considered to be an effective factor to out-of-protospacer edits. The effective factor analysis steps for target-strand edits were the same as for the out-of-protospacer edits.

Annotation of genomic elements. The endogenous dU regions, Cas9-dependent off-target sites and Cas9-independent off-target sites were annotated by homer software (v.4.11) with the hg38 reference genome to annotate the genomic elements information. The enrichment information to the genome background was generated by the homer annotatePeaks command.

Targeted amplicon sequencing data analysis. We first grouped targeted amplicon sequencing FASTQ reads by the unique molecular identifier (UMI) and UMI groups that contained fewer than three reads were discarded. For reads in the same UMI group, we corrected sequencing errors and removed PCR duplications to improve the detection limit by a merge step. In this step, the most frequent amplicon reads were accepted as the consensus reads for subsequent analysis. Then the adapter sequences of consensus reads were removed with cutadapt software (v.1.18). Cleaned reads were mapped to the reference index by BWA MEM (v.0.7.17) with default parameters. Next, we generated mpileup files from mapped BAM files using the samtools mpileup command (v.1.9) with parameters -q 20 -Q 20. Finally, the .mpileup files were converted to .bmat files by Detect-seq tools parse-mpileup commands with default settings.

Public data download and analysis. The ATAC-seq and chromatin immunoprecipitation–sequencing data sets of histone modifications were downloaded from the ENCODE database. All sequencing reads were mapped and processed with the ENCODE Data Standards and Prototype Processing Pipeline (https://www.encodeproject.org/data-standards/). The Digenome-seq raw data set was downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) and analyzed according to the provided method. The accession numbers of downloaded public data sets are available in Supplementary Table 4.

Machine learning model for quantitive prediction. We built a machine learning model to predict the quantitive edits in vivo based on Detect-seq signals. We first selected several features as the independent variables, including Detect-seq mutation count of edited Cs, Detect-seq ratio of edited Cs, motif type of edited Cs, edited Cs index to pRBS, sgRNA alignment mismatch count, gap count at PAM distal side, sgRNA alignment seed region mismatch count, seed region gap count, sgRNA alignment total mismatch count, sgRNA alignment total gap count and sgRNA PAM type. Then the dependent variable was set as mutation ratio of edited Cs by targeted amplicon sequencing. Next, we fitted a Gradient Boost Decision Tree model with XGBoost package (v.0.82) and set parameter n_estimators at 50. The tenfold cross-validation result was performed with default settings to estimate the R^2 of the fitted model.

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

Data availability

All data generated for this paper have been deposited at NCBI GEO and are available under accession numbers GSE151265 and GSE152907. Source data are provided with this paper.

Code availability

Detect-seq tools are available at https://github.com/menghaowei/Detect-seq.

Acknowledgements

We thank W. Wei (Peking University) and J. Hu (Peking University) for discussion, W. Wei together with J. Chen (ShanghaiTech University) for kindly providing related plasmids, and J. Liu (Peking University) for help with experiments. We thank the National Center for Protein Sciences at Peking University in Beijing, China, for assistance with FACS and the Fragment Analyzer. Bioinformatics analysis was performed on the High-Performance Computing Platform of the School of Life Sciences. This work was supported by the National Natural Science Foundation of China (grant nos. 21825701 and 91953201), National Key R&D Program (grant no. 2019YFA0110900) and the Peking University Ge Li and Ning Zhao Education Fund.

Author contributions

Z. Lei, H.M., Z. Lv and C.Y. conceived and guided the research. Z. Lei and M.L. led the development of Detect-seq protocol. H.M. developed the computational pipeline for Detect-seq. H.M., H.W. and H.Z. analyzed all high-throughput sequencing data. Z. Lei and H.M. optimized the targeted amplicon sequencing methodology. Z. Lv conducted cellular experiments and molecular cloning assays. Z. Lei executed Detect-seq experiments. L.L., K.Y., X.Z., Y.Z. and Y.Y. assisted with the experiments. Z. Lei, H.M., Z. Lv and C.Y. wrote the paper.

Competing interests

The authors have filed patent applications on related sequencing technologies.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-021-01172-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-021-01172-w.

Correspondence and requests for materials should be addressed to C.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Peer review Information *Nature Methods* thanks Jia Chen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.