CRISPR DNA base editors with reduced RNA off-target and self-editing activities

Julian Grünewald, Ronghao Zhou, Sowmya Iyer, Caleb A. Lareau, Sara P. Garcia, Martin J. Aryee and J. Keith Joung

Cytosine or adenine base editors (CBEs or ABEs) can introduce specific DNA C-to-T or A-to-G alterations. However, we recently demonstrated that they can also induce transcriptome-wide guide-RNA-independent editing of RNA bases, and created selective curbing of unwanted RNA editing (SECURE)-BE3 variants that have reduced unwanted RNA-editing activity. Here we describe structure-guided engineering of SECURE-ABE variants with reduced off-target RNA-editing activity and comparable on-target DNA-editing activity that are also among the smallest Streptococcus pyogenes Cas9 base editors described to date. We also tested CBES with cytidine deaminases other than APOBEC1 and found that the human APOBEC3A-based CBE induces substantial editing of RNA bases, whereas an enhanced APOBEC3A-based CBE, human activation-induced cytidine deaminase-based CBE, and the Petromyzon marinus cytidine deaminase-based CBE Target-AID induce less editing of RNA. Finally, we found that CBES and ABEs that exhibit RNA off-target editing activity can also self-edit their own transcripts, thereby leading to heterogeneity in base-editor coding sequences.

To engineer SECURE-ABE variants, we first used a protein-truncation strategy to reduce the RNA-recognition capability of the optimized ABEmax fusion. ABEmax harbors a single-chain heterodimer of the wild-type Escherichia coli TRNA-specific adenosine deaminase (TadA) monomer, which deaminates adenosines on tRNA, fused to an engineered E. coli TadA monomer that was modified by directed evolution to deaminate DNA adenosines5,6 (Fig. 1a). Because the wild-type TadA monomer should still be capable of recognizing its tRNA substrate, one can envision that this domain might recruit ABEmax to deaminate RNA adenosines that lie in the same, or a similar, sequence motif to that present in the tRNA. Consistent with this idea, a reanalysis of our previously published RNA sequencing (RNA-seq) data revealed that adenosines edited with the highest efficiencies (80–100%) are embedded in a more extended CUACGAA motif (the bolded A indicating the deaminated adenosine), which contrasts to the shorter UA sequence observed across all edits (Fig. 1b). Importantly, the CUACGAA motif matches the sequence surrounding the adenine deaminated in the tRNA substrate of the wild-type E. coli TadA enzyme (Fig. 1b). Therefore, removing the wild-type TadA domain from ABEmax might reduce its RNA-editing activity, and might not have a dramatic impact on its on-target DNA-editing function (Supplementary Note 1). To test this hypothesis, we generated a smaller ABEmax variant lacking this domain that we refer to as miniABEmax (Fig. 1a).

We used RNA-seq to compare the transcriptome-wide off-target RNA-editing activity of miniABEmax to ABEmax in HEK293T cells. Each of these editors and a nickase Cas9 (nCas9) control were assayed with the following three guide RNAs (gRNAs): two targeted to endogenous human gene sites (HEK site 2 and ABE site 16) and one to a site that does not occur in the human genome (non-targeting, NT). We performed these studies in triplicate and sorted for green fluorescent protein (GFP)-positive cells (each editor or nCas9 was expressed as a protein 2A (P2A)–enhanced GFP (EGFP) fusion (Methods)). As an internal control, we confirmed that ABEmax and miniABEmax induced comparable on-target DNA editing with HEK site 2 and ABE site 16 gRNAs (Supplementary Fig. 1a). Edited RNA adenosines were identified from RNA-seq experiments as previously described by filtering out background editing observed with read-count-matched nCas9 negative controls (Methods). Surprisingly, the total number of edited adenosines induced with miniABEmax expression was not consistently lower than what we observed with ABEmax—the two editors induced on average 80-fold and 54-fold more edited adenosines relative to background (determined with a GFP-only negative control) (Fig. 1c and Supplementary Table 1). However, the overall distribution of individual RNA adenine-editing efficiencies induced by miniABEmax were generally shifted to somewhat lower values (Fig. 1d and Supplementary Fig. 1b). In addition, the sequence logos of adenosines (stratified by editing efficiencies) edited by miniABEmax only yielded shorter GU or UA motifs, in contrast to the more extended CUACGAA motif observed with ABEmax (Supplementary Figs. 2a,b).

We reasoned that we might further reduce the off-target RNA-editing activity of miniABEmax by altering amino acid residues within its remaining engineered E. coli TadA domain that could potentially mediate RNA recognition. However, although a crystal structure of isolated E. coli TadA has previously been solved (PDB accession 1Z3A; Fig. 1e), no structural information was available to delineate how this protein might recognize its RNA substrate. To overcome this, we exploited the availability of a S. aureus TadA–tRNA co-crystal structure (PDB accession 2B3J) (Fig. 1e; Methods). Although E. coli and S. aureus TadA have only partial amino acid sequence homology (39.5% identity; data not shown), these two proteins have a high degree of structural homology (Fig. 1e). This similarity enabled us to overlay the two structures and thereby to infer 26 amino acid residue positions in E. coli TadA that are likely to lie near the enzymatic pocket around the substrate tRNA (Fig. 1e). In addition, we mutated three positively charged residues (R13, K20 and R21) in TadA* that we hypothesized might make contacts to
the phosphate backbone of a nucleic acid molecule. We reasoned that reducing the potentially non-specific affinity of miniABEmax in this way might preferentially reduce its Cas9-independent RNA-editing activity, while preserving its Cas9-assisted on-target DNA-editing activity.

We generated 34 miniABEmax variants bearing various substitutions at the amino acid positions described above and screened each editor for on-target DNA-editing and off-target RNA-editing activities in HEK293T cells. To assess on-target DNA editing, we examined the efficiencies of A-to-G edits induced with four gRNAs targeted to different endogenous gene sequences and found that 23 of the 34 variants induced editing comparable to that observed with miniABEmax and ABEmax (Fig. 1f). To screen for off-target RNA-editing activity (using standard transfection conditions, that is, without sorting for GFP expression; Methods), we quantified editing by each of the 34 variants at six RNA adenines previously identified as being highly edited after ABEmax overexpression in HEK293T cells.1 Fourteen of the 34 variants showed reduced editing activity on at least three of the six RNA adenines that we examined as compared to miniABEmax (Fig. 1f). Based on their DNA- and RNA-editing profiles, we chose to carry forward two miniABEmax variants (K20A/R21A and V82G) for more extensive characterization.

To characterize the transcriptome-wide off-target RNA-editing profiles of miniABEmax(K20A/R21A) and miniABEmax(V82G), we performed RNA-seq with each of these variants and the HEK site 2, ABE site 16 and NT gRNAs. In contrast to what we observed with miniABEmax, the K20A/R21A and V82G variants both induced substantially reduced numbers of edited adenines relative to ABEmax but still approximately fourfold and threefold higher numbers, respectively, than background (determined with the GFP-only negative control) (Fig. 1c and Supplementary Table 1). In addition, the distribution of individual RNA-adenine-editing efficiencies for the two variants was shifted predominantly lower with both variants relative to ABEmax and miniABEmax (Fig. 1d and Supplementary Fig. 1b). The sequence logos of the edited RNA adenines that we derived from these experiments showed that miniABEmax(K20A/R21A) and miniABEmax(V82G) maintained a UA motif (Supplementary Fig. 2c).

To more fully characterize the on-target editing efficiencies of miniABEmax(K20A/R21A) and miniABEmax(V82G), we tested each variant (without sorting cells) in a variety of different sequence contexts with gRNAs for 22 genomic sites in HEK293T cells1, miniABEmax(K20A/R21A) and miniABEmax(V82G) retained efficient absolute on-target modification activity (ranges of mean efficiency of 7.9–70.9% and 10.6–59.4%, respectively; Fig. 2a); however, these efficiencies were typically reduced as compared to ABEmax, with relative activity across the 22 sites ranging from 38.8 to 85.5% and 44.3 to 121.3% for the most highly edited base in the editing window for miniABEmax(K20A/R21A) and miniABEmax(V82G), respectively (Fig. 2b). The relative activity reductions with the variants may be more apparent here because of the higher on-target editing activity achieved as compared to our earlier screening results (Fig. 1f), presumably owing to higher transfection efficiencies achieved with a change in the protocol used (Methods). Neither of the variants showed an apparent preference for a particular sequence context adjacent to the edited adenines (Fig. 2a).

Our analysis of ABE activity with 22 gRNAs also identified a new and unexpected imprecise C-to-G base-editing activity within the editing windows of some DNA on-target sites. This C-to-G on-target DNA editing was observed with ABEmax and miniABEmax(V82G) using the HEK site 2, ABE site 7, and FANCF site 1 gRNAs (Supplementary Fig. 3a). This unwanted editing was consistent across replicates, reached frequencies as high as 14.6% with the FANCF site 1 gRNA and was not observed with the nCas9 control (Supplementary Fig. 3a). Interestingly, for all three sites, the C showing this unexpected editing was present at position 6 of the spacer and was preceded by a T at FANCF site 1 and by an A at HEK site 2 and ABE site 7 (Supplementary Fig. 3a). Notably, for FANCF site 1, consistent C-to-T and C-to-A edits were also observed at position C6 (Supplementary Fig. 3b,c). Additional studies will be needed to clarify the mechanism by which ABEs can induce this new type of imprecise base edit and to define the positions and sequence contexts that dictate whether a C within the editing window is subject to this alteration.

We also sought to compare the off-target DNA-editing activity of miniABEmax(K20A/R21A) and miniABEmax(V82G) with that of ABEmax. To do this, we used targeted amplicon sequencing to quantify editing events at ten previously defined potential off-target sites of three gRNAs (targeted to HEK site 2, HEK site 3 and FANCF site 4)1,2. We found that ABEmax and miniABEmax(K20A/R21A) induced comparable editing patterns and efficiencies at all

---

Fig. 1 | Engineering of SECURE-ABE variants with reduced off-target RNA-editing activity. a, Schematic of ABEmax and miniABEmax architectures and an overview of experimental testing of miniABEmax for on-target DNA and off-target RNA editing. Top, light-blue boxes indicate bipartite nuclear localization signals (bpNLSs) at the N and C termi, TadA* indicates mutant TadA 710 (ref. 5) and small gray boxes indicate the 32-amino-acid linkers. Bottom, the green shape indicates the nCas9 (in this instance, S. pyogenes Cas9 (SpCas9) with a D10A mutation), the blue and red circles indicate the TadA wild-type and mutant monomers, respectively, and the green circles indicate sites of potential adenine deamination on DNA and RNA. b, An unstratified sequence logo (left) and stratified sequence logos for RNA adenines edited with high (80–100%), middle (50–80%) and low (0–50%) efficiencies (middle right) and mutant monomers, respectively, and the green circles indicate sites of potential adenine deamination on DNA and RNA. b, An unstratified sequence logo (left) and stratified sequence logos for RNA adenines edited with high (80–100%), middle (50–80%) and low (0–50%) efficiencies (middle right) and mutant monomers, respectively, and the green circles indicate sites of potential adenine deamination on DNA and RNA.
ten potential off-target sites (including no detectable mutations on some sites) (Supplementary Fig. 4). miniABEmax(V82G) also exhibited comparable editing efficiencies to ABEmax for eight of the ten potential off-target sites examined, but did induce some consistent, but very low efficiency, edits (range of 0.14–0.21%) at two sites, both of which are potential off-target sites for the HEK site 3 gRNA (Supplementary Fig. 4). Although additional experiments will be required to more fully define the genome-wide off-target profiles of miniABEmax(K20A/R21A) and miniABEmax(V82G), these initial studies suggest that the two variants do not exhibit dramatic alterations in their off-target DNA-mutation activity relative to ABEmax.

Having previously shown that off-target RNA editing occurs with a CBE harboring the rAPOBEC1 enzyme (BE3) 5, we wanted to determine whether CBEs harboring other cytidine deaminases such as human APOBEC3A (hA3A)13, enhanced A3A (eaA3A; an engineered A3A with more precise and specific DNA-editing activity), human activation-induced cytidine deaminase (hAID)7 or a sea lamprey (Petromyzon marinus) cytidine deaminase CDA1 (pmCDA1)4 might also induce unwanted edits. To do this, we transfected HEK293T cells in triplicate with plasmids expressing each of these CBEs and a gRNA targeting a site in the RNF2 gene. We then sorted cells with high CBE expression (top 5% of GFP signal) for isolation of genomic DNA (for on-target DNA amplicon sequencing) and total RNA (for RNA-seq) (Methods). At the RNF2 on-target site, hA3A-BE3, eaA3A-BE3, and hAID-BE3 induced mean editing efficiencies of 91%, 82% and 32%, respectively, at position C6, and Target-AID (with a pmCDA1 deaminase at its C-terminal end) showed a mean editing efficiency of

---

**Diagram a**

- ABEmax
- miniABEmax
- SpCas9 (D10A)
- ABEmax
- miniABEmax
- K20A/R21A
- miniABEmax
- V82G

**Diagram b**

- DNA
- RNA
- A-to-I editing (%)

**Diagram c**

- HEK site 2
- ABE site 16
- NT

**Diagram d**

- Number of RNA edits
- RNA
- A-to-I editing (%)

**Diagram e**

- E. coli TadA
- S. aureus TadA + tRNA
- Overlay of E. coli TadA and S. aureus TadA + tRNA
- E. coli TadA + tRNA with highlighted candidate residues of interest

**Diagram f**

- nCas9 Control
- ABEmax
- miniABEmax
- R13A
- K20A
- R21A
- K20A/R21A
- R23A
- E25A
- R26A
- E27A
- V28G
- V30G
- V30W
- N46A
- A48G
- I49A
- A56G
- V82G
- V82W
- E85A
- P86A
- C87A
- C90A
- V106G
- N108A
- A109G
- T111A
- A138G
- A142G
- A142G/A143G
- F148A
- F148A/F149A
- P152A
- V155G
- V155W

**Table**

| Bit Range     | Number of RNA edits | RNA A-to-I editing (%) |
|---------------|---------------------|------------------------|
| 0–100%        | 37,059              | |
87.1% at position C3 (Fig. 3a). RNA-seq experiments revealed that hA3A-BE3 induced tens of thousands of C-to-U edits (Fig. 3b and Supplementary Table 1) distributed throughout the transcriptome (Supplementary Fig. 5a). A number of these Cs were edited with very high (>80%) efficiencies (Supplementary Fig. 5b). Sequence logos derived from all Cs edited by hA3A-BE3 showed a consensus UC motif (Supplementary Fig. 5a). However, sequence logos from subsets of Cs stratified by editing efficiencies revealed a more extended consensus sequence of CCAUCR (the bolded C indicating the deaminated cytosine) for those Cs edited at higher efficiencies (Supplementary Fig. 5a), a motif that is consistent with a previous study that characterized RNA cytidines edited by the hA3A enzyme. By contrast, eA3A-BE3 showed a dramatically reduced number of RNA edits relative to hA3A-BE3 but still slightly more (average of approximately threefold) than the observed background in the GFP-only negative control (Fig. 3b, Supplementary Fig. 5b and Supplementary Table 1). Interestingly, hAID-BE3 and Target-AID induced numbers of RNA C-to-U edits that were comparable to what was observed in the negative control (Fig. 3b, Supplementary Fig. 5b and Supplementary Table 1). The absence of detectable RNA editing in the hAID-BE3 experiments is consistent with a previous study that showed overexpression of an isolated AID enzyme in activated B cells did not yield evidence for RNA editing. By comparison, our two previously described SECURE-BE3 variants induced numbers of RNA C-to-U edits that were slightly higher (BE3(R33A)) than eA3A-BE3, hAID-BE3 and Target-AID, or comparable to the observed background (BE3(R33A/K34A)) (Fig. 3b).

Given their abilities to edit the endogenous human cell transcriptome, we wondered whether CBEs and ABEs might also self-edit their own transcripts, thereby potentially generating sets of heterogeneous base-editor proteins. To assess this, we used our analysis pipeline to quantify self-editing events in our previously published RNA-seq data performed with BE3 expressed at standard levels or overexpressed in HEK293T cells. We observed C-to-U edits at 83–125 and 149–177 different C positions distributed throughout the BE3 transcript with standard expression and overexpression of BE3, respectively (Fig. 4a,b, Supplementary Fig. 6a,b and Supplementary Table 2); efficiencies of C-to-U editing among replicates ranged from 7.3% to 30.4% with standard BE3 expression.
and from 7.1% to 46% with overexpression. Absolute numbers of missense mutations created by these edits ranged from 25 to 44 and from 55 to 64 among replicates with BE3 standard expression and overexpression, respectively (Supplementary Table 2). Importantly, even when overexpressed, the two SECURE-BE3 variants (R33A and R33A/K34A) did not induce any detectable C-to-U edits in their own transcripts (Fig. 4b, Supplementary Fig. 6b and Supplementary Table 2). We observed similar results with BE3 and SECURE-BE3 variants expressed in HepG2 cells (Fig. 4b, Supplementary Fig. 6b and Supplementary Table 2). In addition, self-editing was observed with hA3A-BE3 overexpression in HEK293T cells (28–31 cytosine positions edited with efficiencies ranging from 4.5% to 33.4% among the replicates) (Fig. 4c, Supplementary Fig. 6c and Supplementary Table 2). As expected, overexpression of eA3A-BE3, hAID-BE3 and Target-AID in HEK293T cells showed no detectable evidence of self-editing of their respective transcripts (Fig. 4c, Supplementary Fig. 6c and Supplementary Table 2). Similarly, ABEmax and miniABEmax both induced A-to-I changes at dozens (range of 31–68) of positions throughout their own transcripts with editing efficiencies ranging from 7% to 69.8% among replicates performed with three different gRNAs (Fig. 4d, Supplementary Fig. 6d and Supplementary Table 2). Nearly all of the edits induced by the ABEs are expected to induce missense mutations (Supplementary Table 2). On average, 57% of adenine positions self-edited by ABEmax appeared to be edited across all three replicates (Fig. 4c). Comparing the unions of self-edits from different gRNAs showed 65.85% of overlap between edits across the three gRNAs, suggesting that self-editing is independent of the gRNA with which the ABE was co-expressed (Fig. 4f). Notably, the two miniABEmax variants showed substantially reduced self-editing activity. miniABEmax(K20A/R21A) induced only small numbers (range of 1–3) of self-edits and miniABEmax(V82G) did not induce any detectable self-edits (Fig. 4d, Supplementary Fig. 6d and Supplementary Table 2).

In light of our observation of self-editing, we wondered whether CBEs and ABEs might also be able to edit gRNAs. Although our RNA-seq experiments used RNA extracted from cells by methods optimized for isolation of fragments >200 bases in length, we were nonetheless able to observe thousands of gRNA reads in each of our sequencing data replicates. Therefore, we used our analysis pipeline (Methods) to assess gRNA edits in our RNA-seq data. We did not detect any C-to-U editing of the gRNAs in RNA-seq experiments performed with any of the various CBEs (BE3, BE3(R33A), BE3(R33A/K34A), hA3A-BE3, eA3A-BE3, hAID-BE3 or Target-AID) (Supplementary Fig. 7a-c). Analysis of RNA-seq data from our ABE experiments revealed reproducible editing of an A that resides in the loop of stem loop 2 of the transactivating crRNA (Supplementary Fig. 7d). Edits at this position were present at frequencies of 4.5–19.9% and were most consistently observed with miniABEmax and miniABEmax(V82G) although edits could also be observed in some replicates with ABEmax and miniABEmax(K20A/R21A) (Supplementary Fig. 7d). Given the location and low frequency of this edit, we would not expect it to have a major impact on either activity or specificity of the gRNA–ABE complex.

The work described here extends our understanding of the off-target RNA-editing activity of DNA base editors, expands the options available to minimize these unwanted effects, and
provides new SECURE base editor architectures with other desirable properties. The successful engineering of SECURE-ABE variants shows that, as we previously found with the BE3 CBE, it is possible to minimize unwanted RNA editing while retaining reasonably efficient on-target DNA editing for an ABE. In addition, our characterization of additional CBEs with deaminases other than APOBEC1...
for suggesting the overlay of S. aureus TadA structures and S.J. Lee for suggestion to delete the wild-type TadA monomer from ABEmax. We thank A. Lapinaite German Research Foundation (DFG). J.K.J. is additionally supported by the Desmond Agency (grant HR0011-17-2-0042). Support was also provided by the National Institutes of Health (grant R01 HG009490 to J.K.J. and J.G.), and grant R35 GM118158 to J.K.J. and M.J.A.). J.G. was supported by a research fellowship (GR 5129/1-1) of the German Research Foundation (DFG). I.K.J. is additionally supported by the Desmond and Ann Heathwood MGH Research Scholar Award. We thank G. Ciaramella for the suggestion to delete the wild-type TadA monomer from ABEmax. We thank A. Lapinaite for suggesting the overlay of E. coli and S. aureus TadA structures and S.J. Lee for technical assistance.

Author contributions
All wet lab experiments were performed by R.Z. and J.G. S.L., C.A.L., S.P.G. and M.J.A. performed computational analysis of the data. I.G. and J.K.J. conceived of and designed the study. J.G., M.J.A. and J.K.J. supervised the work. J.G. and J.K.J. wrote the initial manuscript draft and all authors contributed to the writing of the final manuscript.

References
1. 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).
2. Komor, A. C., Kost, Y. R., 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).
3. Gaudelli, N. M. et al. Programmable base editing of AT to GC in genomic DNA without DNA cleavage. Nature 551, 464–471 (2017).
4. Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science 353, aab7296 (2016).
5. Grünwald, J. et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. Nature 569, 433–437 (2019).
6. Gehrke, J. M. et al. An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. Nat. Biotechnol. 36, 977–982 (2018).
7. Komor, A. C. et al. Improved base excision repair inhibition and bacteriophage Mu immunity in ABEmax. Nature 569, 433–437 (2019).
8. Wolf, J., Gerber, A. P. & Keller, W. AAD, an essential RNA-specific adenosine deaminase from Escherichia coli. EMBO J. 21, 3841–3851 (2002).
9. Koblan, L. W. et al. Improving cytidine and adenosine base editors by expression optimization and ancestral reconstruction. Nat. Biotechnol. 36, 843–848 (2018).
10. Kim, J. et al. Structural and kinetic characterization of Escherichia coli TadA, the wobble-specific tRNA deaminase. Biochemistry 45, 6407–6416 (2006).
11. Losey, H. C., Ruthenburg, A. J. & Verdin, G. L. Crystal structure of Staphylococcus aureus tRNA deaminase TadA in complex with RNA. Nat. Struct. Mol. Biol. 13, 153–159 (2006).
12. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas9. Nat. Biotechnol. 33, 187–197 (2015).
13. Wang, X. et al. Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion. Nat. Biotechnol. 36, 946–949 (2018).
14. Sharma, S., Patnaik, S. K., Kemer, Z. & Baysal, B. E. Transient overexpression of exogenous APOBEC3A causes C-to-U RNA editing of thousands of genes. RNA Biol. 14, 603–610 (2017).
15. Fritz, E. L. et al. A comprehensive analysis of the effects of the deaminase AID on the transcriptome and methylome of activated B cells. Nat. Immunol. 14, 749–755 (2013).
16. Zhou, C. et al. Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. Nature 571, 275–278 (2019).
17. Rees, H. A., Wilson, C., Doman, J. L. & Liu, D. R. Analysis and minimization of cellular RNA editing by DNA adenosine base editors. Sci. Adv. 5, eaax5717 (2019).
Competing interests
J.K.J. has financial interests in Beam Therapeutics, Editas Medicine, Excelsior Genomics, Pairwise Plants, Poseida Therapeutics, Transposagen Biopharmaceuticals and Verve Therapeutics (f/k/a Endcadia). The interests of J.K.J. were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies. M.J.A. holds equity in Excelsior Genomics. J.K.J. is a member of the Board of Directors of the American Society of Gene and Cell Therapy. J.G., R.Z. and J.K.J. are co-inventors on patent applications that have been filed by Partners Healthcare/Massachusetts General Hospital on engineered base editor architectures that reduce RNA-editing activities and increase their precision.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0236-6.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.K.J.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
**Methods**

PyMOL analysis of TadA structures. _E. coli_ TadA (PDB accession 1Z3A) and _S. aureus_ TadA with rRNA (PDB accession 2B3J) structures were downloaded from the PDB and visualized with PyMOL. v2.2.2. Subunit A (monomer) of _S. aureus_ TadA with rRNA was superimposed with subunit A of _E. coli_ TadA using the ‘super’ command. All related illustrations (Fig. 1e) were generated with PyMOL (Schrodinger).

**Plasmid cloning.** All ABE constructs (reported in Supplementary Table 3) were cloned using the backbone and the P2A-EGFP-NLS fragment of ABEmax-P2A-EGFP-NLS (Aeg and Notl digest; Addgene, 12181). ABEmax and variants were expressed under a cytomegalovirus promoter (pCMV). Control experiments were performed with a nCas9 negative control that did not contain any TadA domains. All CBE constructs (reported in Supplementary Table 3) were cloned using the backbone of SOI817 and expressed under a CAG promoter (Aeg, Notl and FokI and Notl digestion). For P2A–EGFP fragments of these constructs, we used BP9K335 (pCMV-BE3-P2A-EGFP) as a template. APOBEC3A constructs were cloned using JMGG377 (pCMV-hA3A BE3) as a template. hAID–BE3 was obtained from Addgene (100803). For all CBE plasmids based on the BE3 architecture, nCas9-UGI-NLS-P2A-EGFP (pUL1001; Addgene, 123611) was used as a negative control. For Target-AID, we used NLS-nCas9-NLS-SH3-3xFLAG-NLS-UGI-P2A-EGFP as a separate negative control. Compared to the reference sequence of pmCDA1 from NCBI (ABO15149.1), the pmCDA1 used in Target-NLS-UGI-P2A-EGFP as a separate negative control. Compared to the reference sequence of pmCDA1 from NCBI (ABO15149.1), the pmCDA1 used in Target-NLS-UGI-P2A-EGFP as a separate negative control. Compared to the reference sequence of pmCDA1 from NCBI (ABO15149.1), the pmCDA1 used in Target-NLS-UGI-P2A-EGFP as a separate negative control. Compared to the reference sequence of pmCDA1 from NCBI (ABO15149.1), the pmCDA1 used in Target-NLS-UGI-P2A-EGFP as a separate negative control. Compared to the reference sequence of pmCDA1 from NCBI (ABO15149.1), the pmCDA1 used in Target-NLS-UGI-P2A-EGFP as a separate negative control.

**DNA extraction.** BE3-transfected control cells from the same day.

**FACS sorting.** For ABE DNA off-target experiments in 6-well plates, cells were washed with PBS, trypsinized, centrifuged and gDNA was extracted with QIAmp DNA Mini Kit (Qiagen).

**RNA extraction and reverse transcription.** Cells were lysed to extract RNA.

**Library preparation for targeted amplicon sequencing of DNA or cDNA.** Next-generation sequencing (NGS) of DNA or cDNA was performed as previously described.

**RNA library preparation and sequencing.** RNA-seq experiments were performed as previously described. In brief, RNA libraries were prepared with the TruSeq Stranded Total RNA Library Prep Gold kit (Illumina) following the manufacturer’s instructions. SuperScript III (Invitrogen) was used for first-strand synthesis and iScript for Illumina TruSeq RNA unique dual indexes (96 indexes) were used to avoid index hopping. The libraries were pooled on the basis of quantitative PCR measurements with the NEBNext Library Quant Kit for Illumina (NEB) and paired-end sequenced (2 × 150) on the Illumina MiSeq machine using 300-cycle MiSeq Reagent Kit v2 or Micro Kit v2 (Illumina). FASTQs (post-demultiplexing) were downloaded from Illumina BaseSpace and analyzed using a batch version of CRISPReSSo2.

**Amplicon sequencing analysis.** Amplicon sequencing data were analyzed with CRISPReSSo2 v2.0.27. The heat maps for the SECURE-ABE screening in Fig. 1f display the highest edited adenine on the on-target (DNA) or off-target (RNA) sites. Editing efficiency values were averaged over quadruplicates, log, transformed with a pseudocount of 1 and normalized to ABEmax. Heat maps showing ABE or CBE on-target DNA editing (Figs. 2a and 3a, and Supplementary Fig. 1a) show an editing window that includes the edited As or Cs, respectively, and a gray background for editing efficiencies smaller than 2%. This background cut-off was relaxed for the heat maps showing ABE-induced C-to-N DNA on-target editing (Supplementary Fig. 3) and DNA off-target editing (Supplementary Fig. 4).

**RNA variant-calling pipeline.** All bioinformatic analysis was performed in concordance with GATK best practices for RNA-seq mutation calling, as we have previously described. In brief, raw sequencing reads were two-pass aligned to the reference hg38 reference genome with STAR1 using parameters to discard multimegaping reads. After PCR duplicate removal and base recalibration, mutations in RNA-seq libraries were called using GATK HaplotypeCaller. RNA edits in CBE and ABE overexpression experiments were identified using a downstream modification of the GATK pipeline output as we have previously described. Specifically, mutation positions called by HaplotypeCaller were further filtered to include only those satisfying the following criteria for reference to the corresponding control experiments: (1) read coverage for a given edit in control experiment should be greater than the 90th percentile of read coverage across all edits in the overexpression experiment; and (2) 99% of reads covering each edit in the control experiment were required to contain the reference allele. Edits were further filtered to exclude those with fewer than ten reads or 0% alternate allele frequencies. A–G edits included A–G edits identified on the positive strand as well as T–C edits identified on the negative strand. For CBE overexpression experiments, C–T edits include C–T edits identified on the positive strand as well as T–A edits identified on the negative strand. For ABE overexpression experiments, C–T edits include C–T edits identified on the positive strand as well as T–A edits identified on the negative strand. For ABE overexpression experiments, C–T edits include C–T edits identified on the positive strand as well as T–A edits identified on the negative strand. For ABE overexpression experiments, C–T edits include C–T edits identified on the positive strand as well as T–A edits identified on the negative strand.
Letters

NATURE BIOTECHNOLOGY

had (1) read coverage of at least 50 in all replicates of control and overexpression experiments; (2) 99% reads in all control experiments containing reference allele; and (3) at least 60% alternate allele frequencies in all replicates. From this list, primers were tested for the top 15 edited sites that were also within 150 bases of an exon–exon junction and the six highest edited sites with robust amplification from cDNA were chosen.

To identify self-edits occurring on the base-editing construct, we generated a modified hg38 reference genome with additional contigs for the gRNA and base editor constructs. These additional contigs were appended to the reference genome and each library was reprocessed using GATK best practices, including variant calling with HaplotypeCaller. Variants were then further filtered using a similar process as described above for the transcriptome (that is, filtering for no more than 1% editing in the negative control) with the exception that positions poorly covered in the control owing to differences in the construct design (that is, the deaminase domain) were not filtered out. We note that as both control and BE constructs were expressed from plasmids, the overall expression of these transcripts is much higher than most detected genes, which supersedes the control of coverage between control and BE expression in this analysis (see transcriptome variant calling above). Editing efficiencies per position were computed on the basis of the abundance of Gs (ABE) or Ts (CBE) over total coverage from bam-readcount estimated on the PCR deduplicated .bam files. Edits were further filtered to exclude those with fewer than 50 reads or 0% alternate allele frequencies. The stringency of our variant-calling pipeline might result in the underestimation of the numbers of CBE- or ABE-induced cellular RNA edits and self-edits of BE and gRNA transcripts.

Statistics and data reporting. No specific statistical tests were used. Statistical values include mean and median RNA-editing efficiencies. Error bars (Fig. 2b) depict the s.d. and were plotted using GraphPad Prism v8.1.2. Sample sizes were not predetermined with statistical methods. Investigators were not blinded to experimental conditions or outcome assessments. Details regarding statistical tests and experimental design can be found also in the Nature Research Reporting Summary that is attached to this article.

Data availability

Plasmids encoding the SECURE-ABE and various CBE constructs shown in this work are available on Addgene under article number 28203996. The RNA-seq data used in this study have been deposited in the Gene Expression Omnibus (GEO) under accession GSE129894. Targeted amplicon sequencing data have been deposited at the Sequence Read Archive BioProject accession number PRJNA553185. All other relevant data are available from the corresponding author on request.

Code availability

The authors will make all previously unreported custom computer code used in this work available upon reasonable request.

References

18. Nishimasu, H. et al. Engineered CRISPR–Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).
19. Laird, P. W. et al. Simplified mammalian DNA isolation procedure. *Nucleic Acids Res.* **19**, 4293 (1991).
20. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
21. Clement, K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).
22. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
23. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
24. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

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

Plasmids encoding the SECURE-ABE and various CBE constructs shown in this work are available on Addgene under article number 28203996. The RNA-seq data used in this study have been deposited in the Gene Expression Omnibus (GEO) under accession GSE129894. Targeted amplicon sequencing data have been deposited at the Sequence Read Archive BioProject accession number PRJNA553185. All other relevant data are available from the corresponding author on request.

Code availability

The authors will make all previously unreported custom computer code used in this work available upon reasonable request.

References

18. Nishimasu, H. et al. Engineered CRISPR–Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).
19. Laird, P. W. et al. Simplified mammalian DNA isolation procedure. *Nucleic Acids Res.* **19**, 4293 (1991).
20. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
21. Clement, K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).
22. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
23. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
24. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
- Clearly defined error bars
  State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection
Next-generation sequencing data was collected with Illumina HiSeq2500 and NovaSeq 6000 (RNA-seq) as well as MiSeq (targeted amplicon sequencing) instruments. FACS data was generated using a BD FACSAria II. For fluorometric assays (Pico Green) we used the Synergy HT microplate reader (BioTek) using Gen5 software.

Data analysis
PyMol version 2.2.2, BD FACSDiva Software version 6.1.3, Microsoft Excel Version 1808 (Build 10730.20304), CRISPRezzo 2, bam-readcount version 0.8.0, Sseqtk version 1.0-r82-dirty, STAR 2.6.0c, GATK 3.8, Picard version 2.7.1, WebLogo 2.8 and 3.6.0, MSigDB database version 6.2; GSEA/MSigDB website version 6.3, biomaRt version 2.38.0, Python version 2.7, R version 3.5, GraphPad Prism 8.1.2 (227), Gen5 1.11.5.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data. All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All RNA-seq data have been deposited in the GEO data repository via the accession number GSE129894. Targeted amplicon sequencing data have been deposited at the SRA repository under bioproject accession number PRJNA553185.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size: Sample sizes were determined based on the results of other groups in the field who generate reproducible results with similar setups.

Data exclusions: No data were excluded.

Replication: Independent replicates were performed with different cell passages/batches.

Randomization: Samples were not randomized. Covariates were controlled for by running controls in parallel whenever applicable.

Blinding: Blinding was not performed.

Reporting for specific materials, systems and methods

Materials & experimental systems

- n/a Involved in the study
- Unique biological materials
- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants

Methods

- n/a Involved in the study
- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

- Cell line source(s): source: ATCC; cell lines used: HEK293T cells (ATCC CRL-3216) and HepG2 cells (ATCC HB-8065, experiments from Ref5)
- Authentication: STR profiling by ATCC
- Mycoplasma contamination: Supernatant was analyzed every two weeks using MycoAlert PLUS (Lonza). Both cell lines continuously tested negative. Additionally, we used RNA-seq data analysis to rule out mycoplasma infection a posteriori.
- Commonly misidentified lines (See ICLAC register): HEK293T and HepG2 cell lines are not listed in the ICLAC register (version 9).
### 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 | Cell culture and transfection procedures are described in the online methods. Cells were washed and filtered through a 35μm cell strainer cap before sorting (36-40h after transfection). |
|--------------------|------------------------------------------------------------------------------------------------------------------|
| Instrument         | FACSAria II (BD Biosciences)                                                                                       |
| Software           | BD FACSDiva Software v6.1.3                                                                                       |
| Cell population abundance | Cell population abundances after gating for target populations were similar in different experiments, depending on the cell line. HEK293T cells transfected with plasmids described in the supplement usually were ~ 40-60% GFP+ (of gated population = % parent in BD FACSDiva). HepG2 cells (experiments from Ref5) were usually ~10-20% GFP-positive after transfecting BE constructs. |
| Gating strategy    | Gates were established using untransfected control cells and transfected GFP+ cells. Gates were drawn to collect either all GFP+ cells (all ABE experiments shown, except for Fig. 1b, which is top 5% sorted) or subsets of GFP-expressing cells. For overexpression experiments, cells with top 5% of GFP signal were sorted, after gating for the cell population (~5% of parent). In these top 5% sorting experiments, nCas9 and GFP controls were MFI-matched to the top 5% MFI of BE constructs. Please see the Supplementary Information for gating strategies in different experimental contexts. |

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