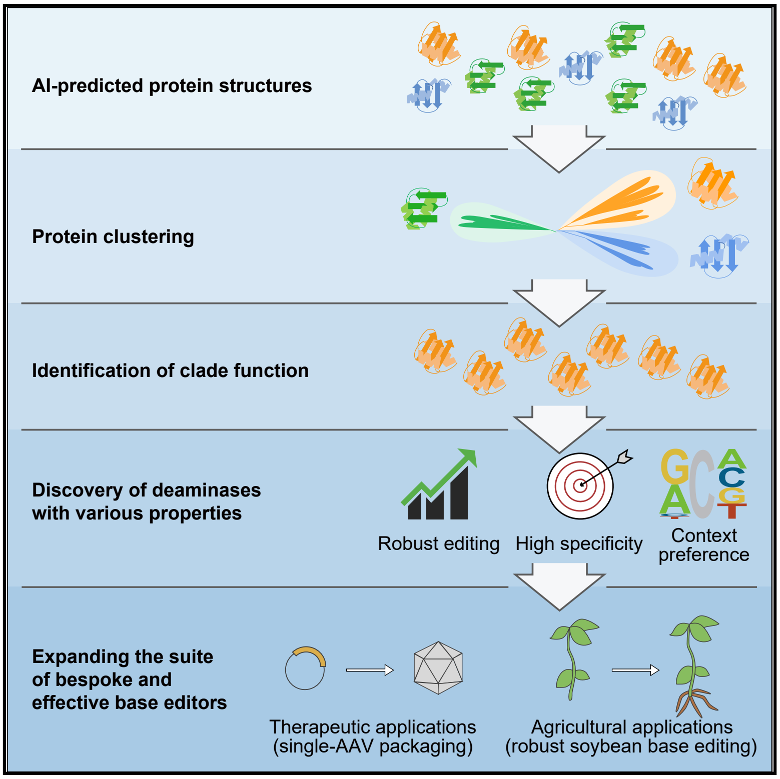
Cell

Article

Discovery of deaminase functions by structure-based protein clustering

Graphical abstract


AI-predicted protein structures
Protein clustering
Identification of clade function
Discovery of deaminases with various properties
Robust editing
High specificity
Context preference
Expanding the suite of bespoke and effective base editors
Therapeutic applications (single-AAV packaging)
Agricultural applications (robust soybean base editing)

Highlights

- AI-guided structural classification establishes new deaminase family relationships
- SCP1.201 deaminase clade contains both ssDNA and dsDNA cytidine deaminases
- Newly deaminases are smaller and have increased activities and minimal off-targets
- Further AI-assisted truncation enables AAV packaging and efficient soybean editing

Authors

Jiaying Huang, Qiupeng Lin, Hongyuan Fei, ..., Jin-Long Qiu, Kevin Tianmeng Zhao, Caixia Gao

Correspondence

kzhao@qi-biodesign.com (K.T.Z.),
cxgao@genetics.ac.cn (C.G.)

In brief

AI-assisted structural predictions and alignments establishes a new protein classification and functional mining method, further discovering a suite of single- and double-stranded deaminases, which show great potential as bespoke base editors for therapeutic or agricultural breeding applications.

Huang et al., 2023, Cell *186*, 1–14
July 20, 2023 © 2023 Elsevier Inc.
https://doi.org/10.1016/j.cell.2023.05.041

CellPress

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

Cell

CellPress

Article

Discovery of deaminase functions by structure-based protein clustering

Jiaying Huang,[1,9] Qiupeng Lin,[1,9] Hongyuan Fei,[1,2,9] Zixin He,[1,2,9] Hu Xu,[3] Yunjia Li,[1,2] Kunli Qu,[4] Peng Han,[4] Qiang Gao,[3] Boshu Li,[1,2] Guanwen Liu,[1] Lixiao Zhang,[3] Jiacheng Hu,[1] Rui Zhang,[1] Erwei Zuo,[5] Yonglun Luo,[4,6] Yidong Ran,[3] Jin-Long Qiu,[7,8] Kevin Tianmeng Zhao,[3,*] and Caixia Gao[1,2,10,*]

[1]State Key Laboratory of Plant Cell and Chromosome Engineering, Center for Genome Editing, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing, China

[2]College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing, China

[3]Qi Biodesign, Beijing, China

[4]Lars Bolund Institute of Regenerative Medicine, Qingdao-Europe Advanced Institute for Life Sciences, BGI-Qingdao, BGI-Shenzhen, Qingdao, China

[5]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

[6]Department of Biomedicine, Aarhus University, 8000 Aarhus, Denmark

[7]State Key Laboratory of Plant Genomics, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

[8]CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing, China

[9]These authors contributed equally

[10]Lead contact

*Correspondence: kzhao@qi-biodesign.com (K.T.Z.), cxgao@genetics.ac.cn (C.G.)

https://doi.org/10.1016/j.cell.2023.05.041

SUMMARY

The elucidation of protein function and its exploitation in bioengineering have greatly advanced the life sciences. Protein mining efforts generally rely on amino acid sequences rather than protein structures. We describe here the use of AlphaFold2 to predict and subsequently cluster an entire protein family based on predicted structure similarities. We selected deaminase proteins to analyze and identified many previously unknown properties. We were surprised to find that most proteins in the DddA-like clade were not double-stranded DNA deaminases. We engineered the smallest single-strand-specific cytidine deaminase, enabling efficient cytosine base editor (CBE) to be packaged into a single adeno-associated virus (AAV). Importantly, we profiled a deaminase from this clade that edits robustly in soybean plants, which previously was inaccessible to CBEs. These discovered deaminases, based on AI-assisted structural predictions, greatly expand the utility of base editors for therapeutic and agricultural applications.

INTRODUCTION

The discovery and engineering of proteins has greatly transformed the life sciences. Traditional enzyme mining, based solely on sequence information, has been effective at classifying and predicting protein functions and evolutionary trajectory.[1,2] However, one-dimensional (1D) information, whether in the form of core amino acids, specific motifs, overall amino acid sequence identity, or hidden Markov models (HMMs), cannot completely illuminate the functional characteristics of proteins.

In contrast, since protein function is ultimately determined by three-dimensional (3D) protein folds, understanding protein structures would provide reliable and rational insights into protein function during the process of protein mining and clustering classifications.[3,4] Although the number of publicly reported protein structures is increasing, it is miniscule compared with the number of proteins discovered based on amino acid sequences.[5,6] Recently, many artificial intelligence (AI) methods have been developed that use 1D amino acid sequences to accurately predict high-resolution 3D protein structures.[7–9] These protein structure prediction methods should thus enable large-scale mining and classifications of proteins with specific functions.

Deaminase-like proteins catalyze the deamination of nucleotides and bases in nucleic acids. They play important roles in defense, mutation, nucleic acid metabolism, and in other biological processes[10–13] and have been recently exploited for use in programmable DNA and RNA base editors,[14–16] a class of precise genome editing technologies. Members of this family act as nucleotide deaminases and nucleic acid deaminases, including adenosine, cytidine, and guanine deaminases, and have the ability to act on single-stranded DNA (ssDNA),[17] double-stranded DNA (dsDNA),[10] double-stranded RNA (dsRNA),[18] transfer RNA (tRNA),[19] free nucleosides,[12] and other nucleotide derivatives.[20] The sporadic distribution of deaminases and their rapid evolution due to positive selection often confound the relationships between the various protein families in phylogenetic analyses based on sequence.[20,21] Here, we performed protein clustering classifications on the greater deaminase family of proteins, based on AlphaFold2-predicted 3D structures.

To better differentiate and discover deaminases with diverse functions, we employed AlphaFold2 to predict deaminase structures and subsequently performed structural comparisons to generate a taxonomic tree of deaminase proteins that better reflects the different types of cytidine deaminases. Using AlphaFold2-predicted structures, we were able to classify proteins into different clades more efficiently than by using 1D amino acid sequences.

Cytosine base editors (CBEs) use cytidine deaminases to catalyze C-to-U base conversions, resulting in permanent C•G-to-T•A base edits in DNA.[14,15,22,23] Base editors have great potential in therapeutic genome editing, in fundamental life sciences research, and for breeding new elite traits into plants.[24–26] Previous DNA base editors exploited the use of two types of cytidine deaminases acting on either ssDNA or dsDNA.[10,14] To date, only a few ssDNA-targeting apolipoprotein B mRNA-editing enzyme catalytic polypeptide (APOBEC)/activation-induced cytidine deaminase (AID)-like deaminases and one dsDNA-targeting deaminase (DddA) have been used to generate CBEs.[10,14,15,27–30] These deaminases remain limited to sequence context restrictions, low on-target:off-target editing ratios, and large protein sizes, which makes their delivery by adeno-associated virus (AAV) viral vectors difficult.[31] For unknown reasons, some species like soybean plants, a staple agricultural crop grown all over the world, have suffered from poor cytosine base editing since the technology was first introduced in 2016.[32] Thus, robust and more efficient CBEs are still needed to further expand their utility. By generating protein classifications based on predicted structures, we have developed a suite of ssDNA deaminases (Sdds) and dsDNA deaminases (Ddds) used for precision genome editing. We highlight that enzyme mining based on structures predicted by AlphaFold2 is a simple, flexible, and high-throughput method to classify and engineer proteins with unknown functions.

RESULTS

Clustering and discovery of new cytidine deaminases via protein structures

We hypothesized that the comparison and clustering of known or predicted protein structures—given that the 3D structure of a protein ultimately determines its function—could be an effective method for classifying deaminases into functional clades. Thus, we employed a combination of AI-assisted protein structure prediction, structural alignments, and clustering to generate protein classification relationships among deaminases (Figure 1A). We selected 238 protein sequences annotated as having a deaminase domain from the InterPro database and 4 distant outgroup candidate protein sequences from the c-Jun activation domain binding protein (JAB)-domain family (Figure S1A; Table S1). Specifically, we randomly selected 15 candidates of at least 100 amino acids in length from each of the 16 deaminase families and used AlphaFold2 to predict their protein structures. We conducted multiple structural alignments (MSTAs) of all candidates, and based on the MSTA results, we generated candidate similarity matrices reflecting the overall structural correlation between the proteins. We then organized these similarity matrices into a structural dendrogram using unweighted pair group method with arithmetic mean (UPGMA)[33] (Figure 1B). The dendrogram clustered the 238 proteins into 20 unique structural clades, and the deaminases within each clade have distinct conserved protein structural domains (Figures 1C and 1D).

We found that accurate protein clustering classifications could be generated based on protein structural alignments, even without the use of contextual information such as conserved gene neighborhoods and domain architectures. When using structure-based hierarchical clustering, different clades reflect unique structures, implying distinct catalytic functions and properties (Figure 1D). Interestingly, we also found that structure-based clustering methods were much more robust and effective at sorting for functional similarities than traditional 1D amino acid sequence-based clustering approaches (Figures S1B and S1C). For example, adenosine deaminases (A_deamin, PF02137 in InterPro database), enzymes involved in purine metabolism, were split into different clades when using amino acid sequence-based clustering methods but were all grouped together into a single A_deamin clade using our structure-based clustering approach (Figures 1B, 1C, and S1B). Additionally, four deaminase families (deoxycytidylate monophosphate [dCMP], MafB19, LmjF365940, and APOBEC, as annotated by InterPro) were each divided into two separate clades when using structure-based clustering (Figures 1C and 1D). Comparison of protein structures showed that the two clades for each of these four deaminase families had quite different structures, contrary to what their InterPro naming and sequence-based classification might suggest (Figures 1D and S1D–S1H). In summary, AI-assisted 3D protein structures provide reliable clustering results and only require an amino acid sequence from the user, making it a convenient and effective strategy for generating protein relationships.

Evaluating diverse deaminase clades by fluorescence imaging

CRISPR-based CBEs are precise genome editing technologies capable of generating C•G-to-T•A substitutions in the genome of living cells. Because ssDNA-specific cytidine deaminases are an essential component of CBEs, we sought to explore the deamination activity of each structure-based classified deaminase clade in the context of DNA base editing. We evaluated a total of 239 deaminase domains by selecting at least 5 proteins from each clade. Importantly, because the core deaminase domain used for clustering may not show editing activity, we extended each deaminase sequence to include additional secondary structures from each corresponding gene around the deaminase domain (Figure S1A). For each of the 239 newly annotated deaminases, we generated plant CBEs by fusing each candidate domain-related sequence to the N terminus of a Cas9 nickase (nCas9, D10A), followed by an uracil-DNA glycosylase inhibitor (UGI).[14,34] We developed four blue fluorescent protein (BFP)-to-green fluorescent protein (GFP) reporter systems to reflect T<u>C</u>, C<u>C</u>, G<u>C</u>, and A<u>C</u> 5′-base deamination preferences

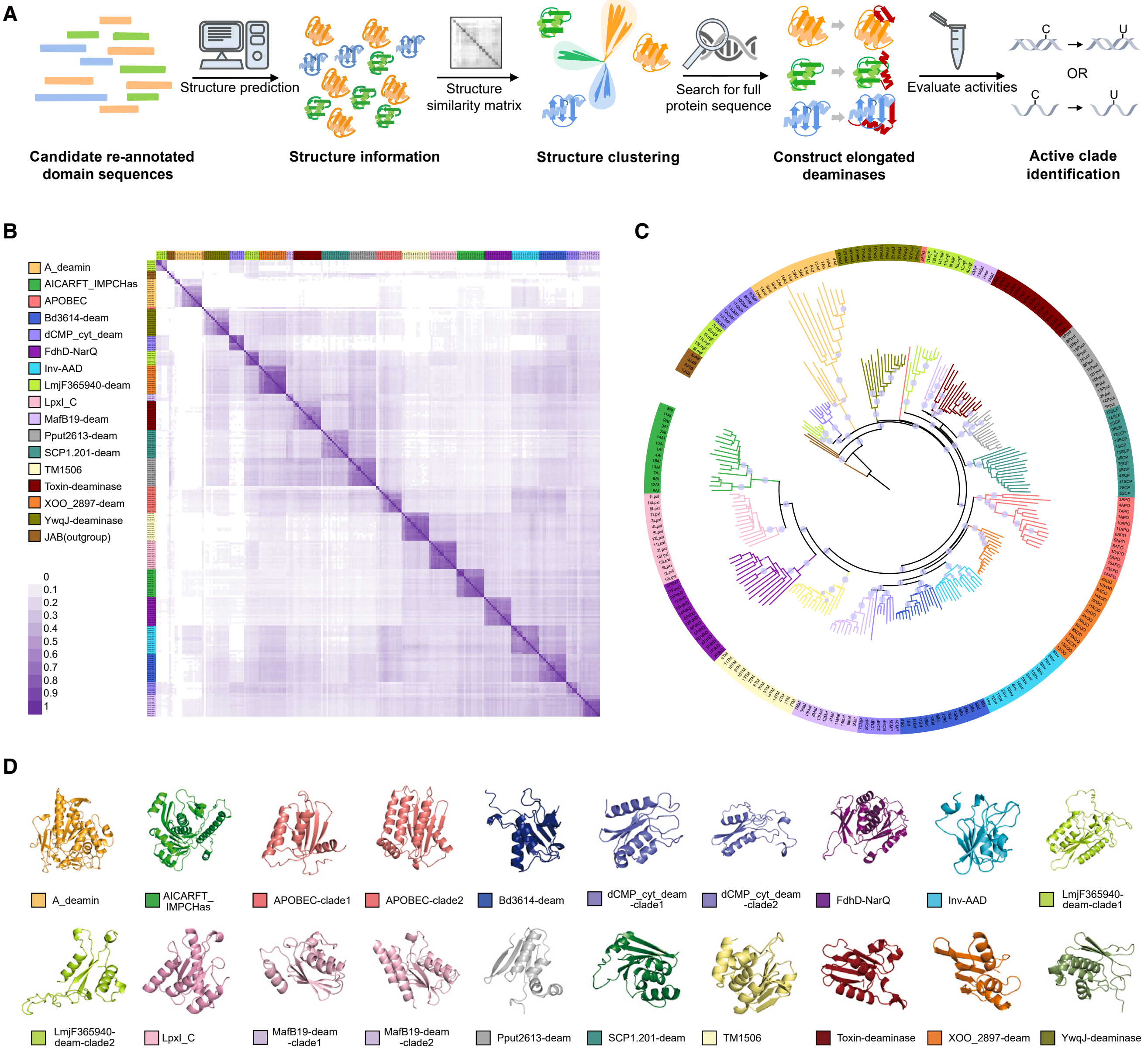


Figure 1. Protein clustering of deaminases based on structures predicted by AlphaFold2

(A) Workflow of protein clustering based on AlphaFold2-predicted structures. The structures of candidate re-annotated domain sequences were predicted by AlphaFold2 and subsequently clustered based on structural similarities. Then, ssDNA and dsDNA cytidine deamination activities were experimentally tested in plant and human cells.

(B) Structural similarity matrix to reflect similarities between 242 predicted protein (238 cytidine deaminases and 4 JAB) structures across 16 deaminase families and 1 outgroup. Different family proteins are distinguished by different colors; heatmap color shades indicate the degree of similarity.

(C) The classification of proteins into different deaminase families based on protein structure and labeled with different color modes. Nodes with Bootstrap ≥90 are identified by circles.

(D) Representative predicted structures for each of 16 deaminase clades.

(Figure S2A). Each CBE was co-transformed with all four BFP-to-GFP reporter plasmids into rice protoplasts and analyzed by fluorescent microscopy after 3 days.[34] We found that deaminases belonging to the SCP1.201 (PF14428), XOO_2897 (PF14440), MafB19 (PF14437), toxin-deaminase (PF14424), and TM1506 (PF08973) clades possessed ssDNA cytidine deamination activity (Table S2). Interestingly, we noticed that some deaminase candidates displayed different sequence preferences, compared with the APOBEC/AID-like deaminases, as evaluated using the fluorescence reporter system (Table S2). Therefore, we demonstrated that the use of 3D structures for protein classification enables the discovery of new functional deaminase clusters for use in base editors, offering new opportunities for developing enhanced and bespoke precise base editing tools.

Validation of the diverse functions of SCP1.201 deaminases

While evaluating deaminases from each clade, we were surprised to find that some deaminases annotated from the SCP1.201 clade were capable of deaminating ssDNA substrates (Table S2). These deaminases were previously named dsDNA deaminase toxin A-like (DddA-like) deaminases in the InterPro database (PF14428). The DddA-like deaminase was recently developed into a CRISPR-free dsDNA CBE (DdCBE) capable of deaminating cytosine bases on dsDNA.[10] Because of DddA, all proteins in the SCP1.201 clade were also annotated as Ddds. To re-analyze this SCP1.201 clade, we selected all 489 SCP1.201 deaminases from the InterPro database. We also included 7 additional proteins that were 35%–50% identical by basic local alignment search tool (BLAST) with DddA but were characterized separately in InterPro. After identity and coverage filtering, we performed a new AI-assisted protein structure-based classification of 332 SCP1.201 deaminases (Table S3). Structure clustering showed that the SCP1.201 deaminases clustered into different clades with unique core structural motifs (Figures 2A–2E and S2B).

We found that DddA and 10 other proteins clustered into one subclade of SCP1.201. Upon analyzing the 3D predicted structures of all 11 proteins within this subclade, we found that they shared a similar core structure to DddA. Given their structural similarities to DddA, we hypothesized that the other proteins in this subclade can also perform dsDNA cytidine deamination. To evaluate dsDNA deamination, we generated DdCBEs comprising each deaminase alone or split in half at a residue similar to the site where DddA was split by protein structure alignment and joined together using a dual transcription activator-like effector (TALE) system[10] (Figure S2A; Table S2). We evaluated 10 proteins from this Ddd subclade in HEK293T cells at the *JAK2* and *SIRT6* sites and observed that 8 proteins could perform dsDNA base editing (Figures 2A and 2F). We hereafter referred to these deaminases as Ddds and assigned them to this newly identified Ddd subclade.

To evaluate other SCP1.201 candidate proteins, we selected at random 76 proteins that are representative of each node branch based on the SCP1.201 structural clustering results (Figure 2A; Table S2) and subjected these to our CBE fluorescent reporter system. We found that 45 showed detectable fluorescence and selected 23 to evaluate endogenous base editing in the context of CBE in mammalian cells (Figure 2; Table S2). Although these were previously characterized as DddA-like, many showed cytosine base editing activity on ssDNA (Figures 2A and 2G; Tables S2 and S4) but not dsDNA (Figures S2C and S2D). Therefore, we hereafter referred to these ssDNA-targeting protein domains from the SCP1.201 clade as Sdds. We were surprised to find that a majority of protein members from the SCP1.201 clade were found to be Sdd proteins, since these were all previously annotated as DddA-like (Tables S2 and S4). We also observed that these Sdd proteins shared a similar protein structure as Sdd7, one of the highest editing ssDNA CBEs, which is distinct from the Ddd proteins (Figures 2B–2E and S2B). Thus, the annotated DddA-like deaminases in the InterPro database (PF14428) should be further subdivided and re-annotated accordingly.

In comparison, we also performed a clustering of the proteins from the SCP1.201 clade, based on 1D amino acid sequences, and found that some outgroup members were dispersed throughout the tree, even though we chose four more closely related families as outgroups (Figures S2E and S2F). These results highlight the usefulness and importance of using protein structure-based classifications for comparing and evaluating protein functional relationships.

New Ddd proteins have distinct editing preferences to DddA

Due to the strict 5′-TC sequence motif preference of DddA, the use of DddA-based dsDNA base editors is limited predominantly to TC targets.[10] Although the recently evolved DddA11 displayed a broadened ability to deaminate and edit cytosine bases with a 5′-HC (H = A, C, or T) motif, the editing efficiency for AC, CC, and GC targets still needs to be improved.[35] We evaluated the newly discovered Ddd proteins to determine if they could expand the utility and targeting scope of DdCBEs. Thus, 13 deaminases belonging to the Ddd subclade were cloned into DdCBEs and evaluated for dsDNA base editing at the endogenous *JAK2* and *SIRT6* sites in HEK293T cells (Figures 2F, S3A, and S3B; Table S2). Interestingly, we found that Ddd1, Ddd7, Ddd8, and Ddd9 have editing efficiencies comparable to or higher than DddA (Figures 3A, S3A, and S3B). Importantly, we identified that Ddd1 and Ddd9 have a much higher editing activity compared with DddA at 5′-GC motifs (Figures 3A, S3A, and S3B). Strikingly, at the C$_{10}$ (5′-GC) residue in *JAK2* and the C$_{11}$ (5′-GC) residue in *SIRT6*, we found that while DddA resulted in 21.1% and 0.6% editing, respectively, Ddd9 was capable of editing 65.7% and 45.7% (Figure 3A).

Because certain Ddd proteins seemed to exhibit distinct editing patterns, compared with DddA, we sought to evaluate any sequence motif preference for these Ddd proteins. We first constructed 16 plasmids[35] encoding the *JAK2* target sequence and modified positions 9–11 from GCC to NCN (N = A, T, C, and G), yielding 16 different plasmids, and we independently co-transfected each plasmid along with a DdCBE variant (Figure 3B). Following comparative analyses of C•G-to-T•A base conversion frequencies for each NCN, we generated corresponding sequence motif logos to reflect sequence context preferences of each Ddd (Figure 3B). We found that as previously discussed, DddA and its structural homolog, Ddd7, strongly preferred a 5′-TC sequence motif (Figures 3C and S3C). In contrast, we found that Ddd1 and Ddd9 showed a preference for editing 5′-GC substrates, while Ddd8 showed a preference for editing 5′-WC (W = A or T) substrates (Figures 3C and S3C). Therefore, the newly discovered dsDNA-targeting deaminases can edit cytosine bases at motifs previous inaccessible to DddA, which is also essential for future engineering efforts.

Sdds enable base editing in human cells and plants

We next wondered whether the newly characterized Sdd proteins could be used for more precise or efficient base editing.

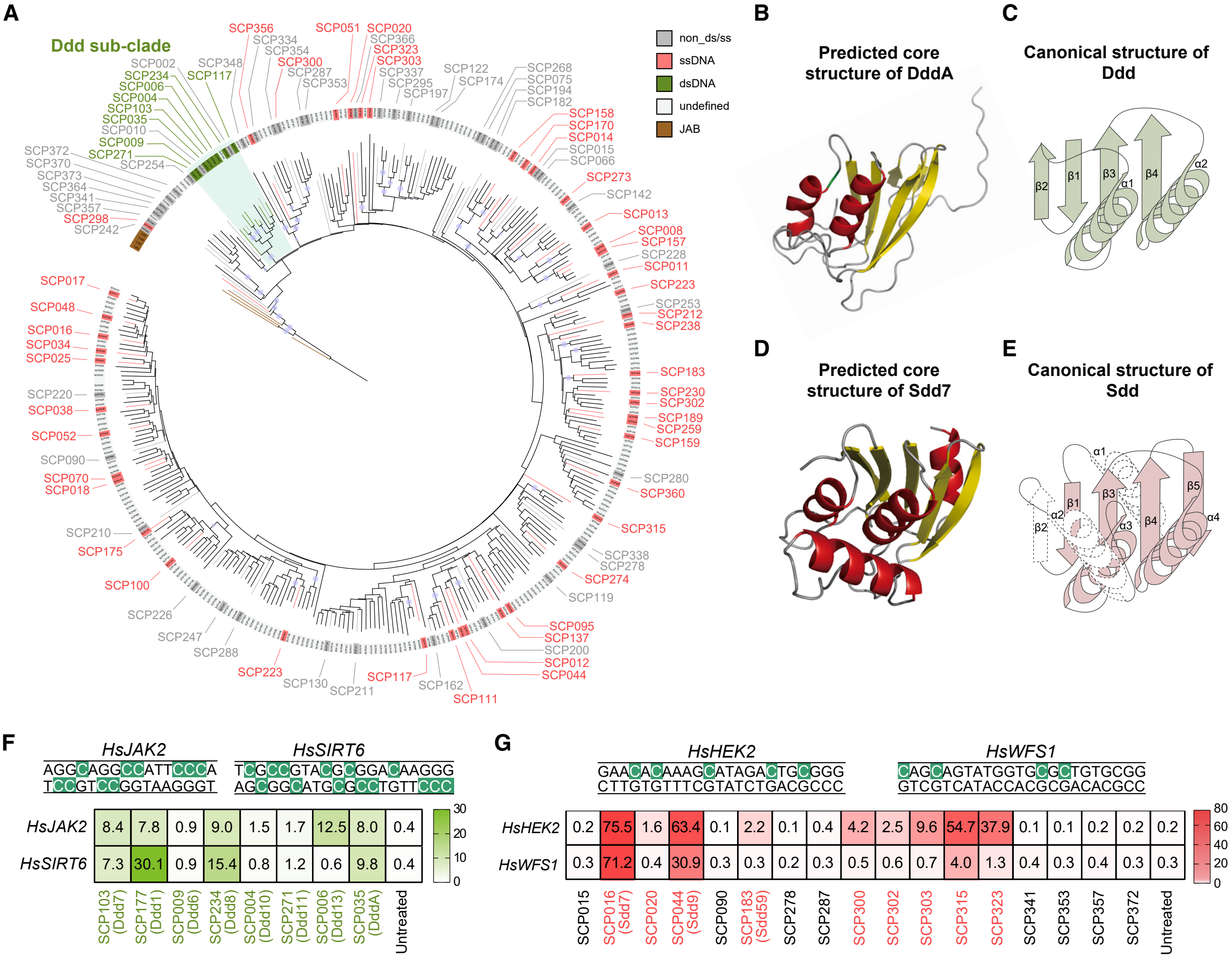


Figure 2. The clustering and characteristics of SCP1.201 deaminases

(A) Classification of SCP1.201 deaminases based on protein structure. The JAB families are colored brown and regarded as an outgroup, and the tested deaminases are shown in red (single-stranded editing), green (double-stranded editing), gray (no editing was detected). Undefined deaminases in white await further functional analysis. The detailed single- and double-stranded activity can be found in Table S4. Nodes with Bootstrap ≥90 are identified by circles.

(B) Predicted core structure of DddA by AlphaFold2.

(C) Characteristics of the canonical structure of Ddd protein.

(D) Predicted core structure of Sdd7 by AlphaFold2.

(E) Characteristics of the canonical structure of Sdd protein.

(F) Experimental evaluation of dsDNA deamination activity of Ddds at two endogenous sites in HEK293T cells. The edited bases used for calculating editing are highlighted in green.

(G) Experimental evaluation of ssDNA deamination activity of Sdds at two endogenous sites in HEK293T cells. The edited bases used for calculating editing are highlighted in green.

Data in (F) and (G) are representative of three independent biological replicates (n = 3).

We chose to evaluate the six most active Sdds as well as four weaker Sdds and compared their activities using a fluorescent reporter system (Table S2). We generated plant CBEs for each of the 10 Sdds and evaluated their endogenous base editing across 6 sites in rice protoplasts (Figures 4A and S4A). We found that seven of the deaminases (Sdd7, Sdd9, Sdd5, Sdd6, Sdd4, Sdd76, and Sdd10) had higher activity, compared with the rat APOBEC1 (rAPOBEC1)-based CBE. The most active Sdd7 base editor reached as high as 55.6% cytosine base editing, which was more than 3.5-fold higher than that of rAPOBEC1. To examine the versatility of these deaminases, we also constructed the corresponding human-cell-targeting BE4max vectors[36] and evaluated their editing efficiencies across three endogenous target sites in HEK293T cells. In agreement with the results in rice, we found that Sdd7 had the highest editing activity (Figure S4B).

We previously showed that human APOBEC3A (hA3A) performed robust base editing with a large editing window in

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

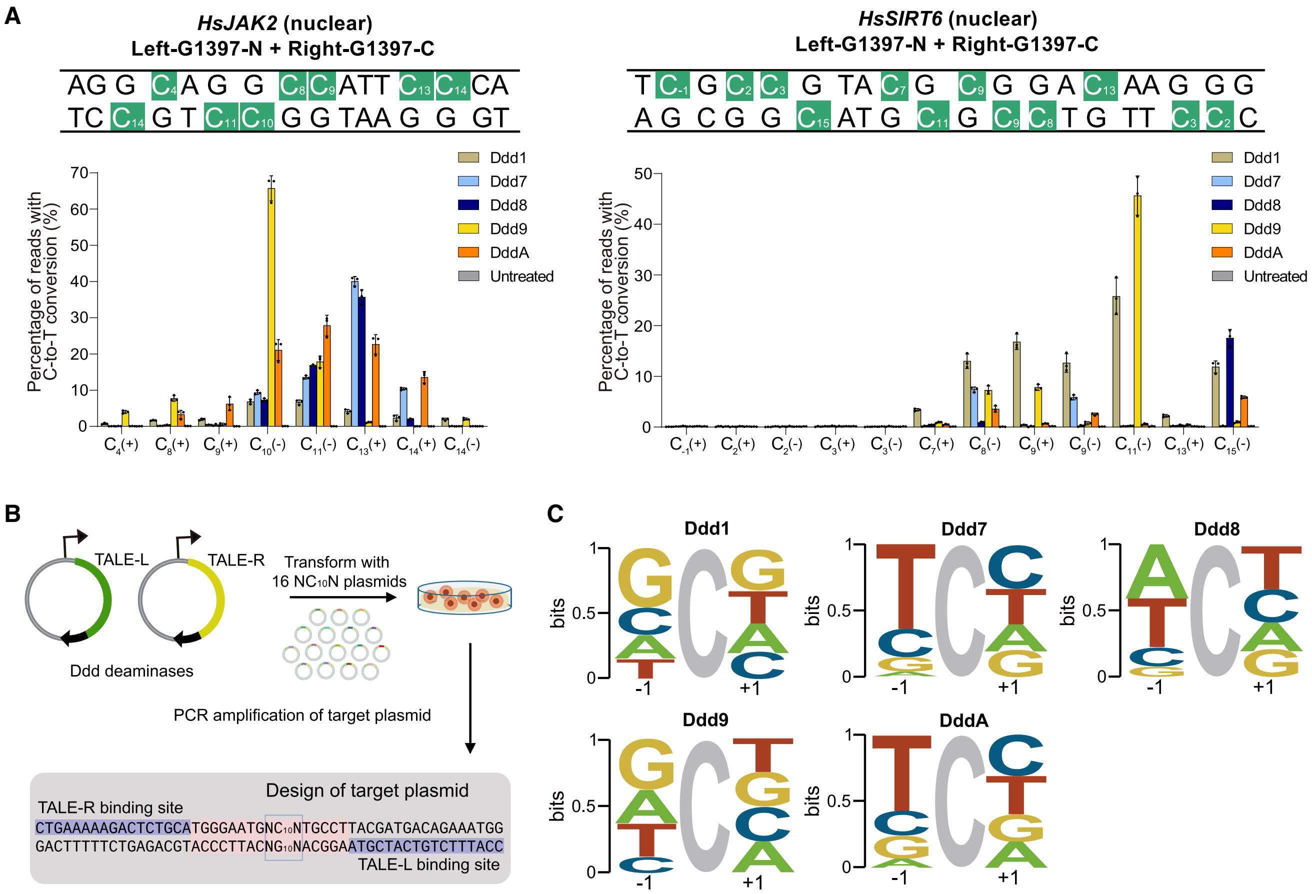


Figure 3. Evaluating newly discovered Ddd protein properties for use as base editors

(A) Editing efficiencies and editing windows of Ddd1, Ddd7, Ddd8, Ddd9, and DddA SCP1.201 dsDNA deaminases at two genomic target sites in HEK293T cells.

(B) Plasmid library assay to profile context preferences of each Ddd protein in mammalian cells. Candidate proteins target and edit the "NC$_{10}$N" motif.

(C) Sequence motif logos summarizing the context preferences of Ddd1, Ddd7, Ddd8, Ddd9, and DddA, as determined by the plasmid library assay.

For all plots, dots represent individual biological replicates, bars represent mean values, and error bars represent the SD of three independent biological replicates (n = 3).

plants.[37,38] We therefore compared the editing activities of hA3A and Sdd7 in human cells (Figure S4B) and plants (Figure S4C). Interestingly, Sdd7 had comparable editing activities to hA3A across all three target sites in HEK293T cells (Figure S4B) and five endogenous sites in rice protoplasts (Figure S4C). Because editing efficiency is of primary significance for genome editing in plant breeding, these results confirmed that Sdd7 is a robust CBE for use in both plants and human cells.

Sdd proteins have unique base editing characteristics

When evaluating endogenous base editing, we observed different editing patterns by the different Sdd-CBEs across all tested genomic target sites in both human and rice cells. For instance, while Sdd7, Sdd9, and Sdd6 showed no particular motif editing preference, Sdd3 seemed to prefer editing 5′-G$\underline{C}$ and 5′-A$\underline{C}$ motifs and strongly disfavor editing 5′-T$\underline{C}$ and 5′-C$\underline{C}$ motifs (Figure S4D). To better profile the editing patterns of each deaminase, we used targeted reporter anchored positional sequencing (TRAP-seq), a high-throughput approach for parallel quantification of base editing outcomes.[39] A 12K TRAP-seq library comprised of 12,000 TRAP constructs, each containing a unique gRNA expression cassette and the corresponding surrogate target site, was stably integrated into HEK293T cells by lentiviral transduction. Following cell culture and antibody selection, base editors were stably transfected into this 12K-TRAP cell line followed by 10 days of blasticidin selection (Figure 4B). On day 11 post transfection, we extracted the genomic DNA and performed deep amplicon sequencing to evaluate the editing products of each deaminase (Figure 4B). We found that Sdd7 and Sdd6 showed no strong sequence context preference, but rAPOBEC1 had a strong preference for 5′-T$\underline{C}$ and 5′-C$\underline{C}$ bases while disfavoring 5′-G$\underline{C}$ and 5′-A$\underline{C}$ bases (Figure 4C). By contrast, Sdd3 showed an entirely complementary pattern preferring to edit 5′-G$\underline{C}$ and 5′-A$\underline{C}$ bases while showing nearly no activity toward 5′-T$\underline{C}$ and 5′-C$\underline{C}$ bases (Figure 4C). Interestingly, we found that Sdd6 and Sdd3 had different editing windows and preferred to edit positions +1 to +3 in the protospacer, as compared with rAPOBEC1 and Sdd7 (Figure 4C). In conclusion, the newly identified Sdd base editors show unique base editing properties such as increased editing efficiencies, disparate

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

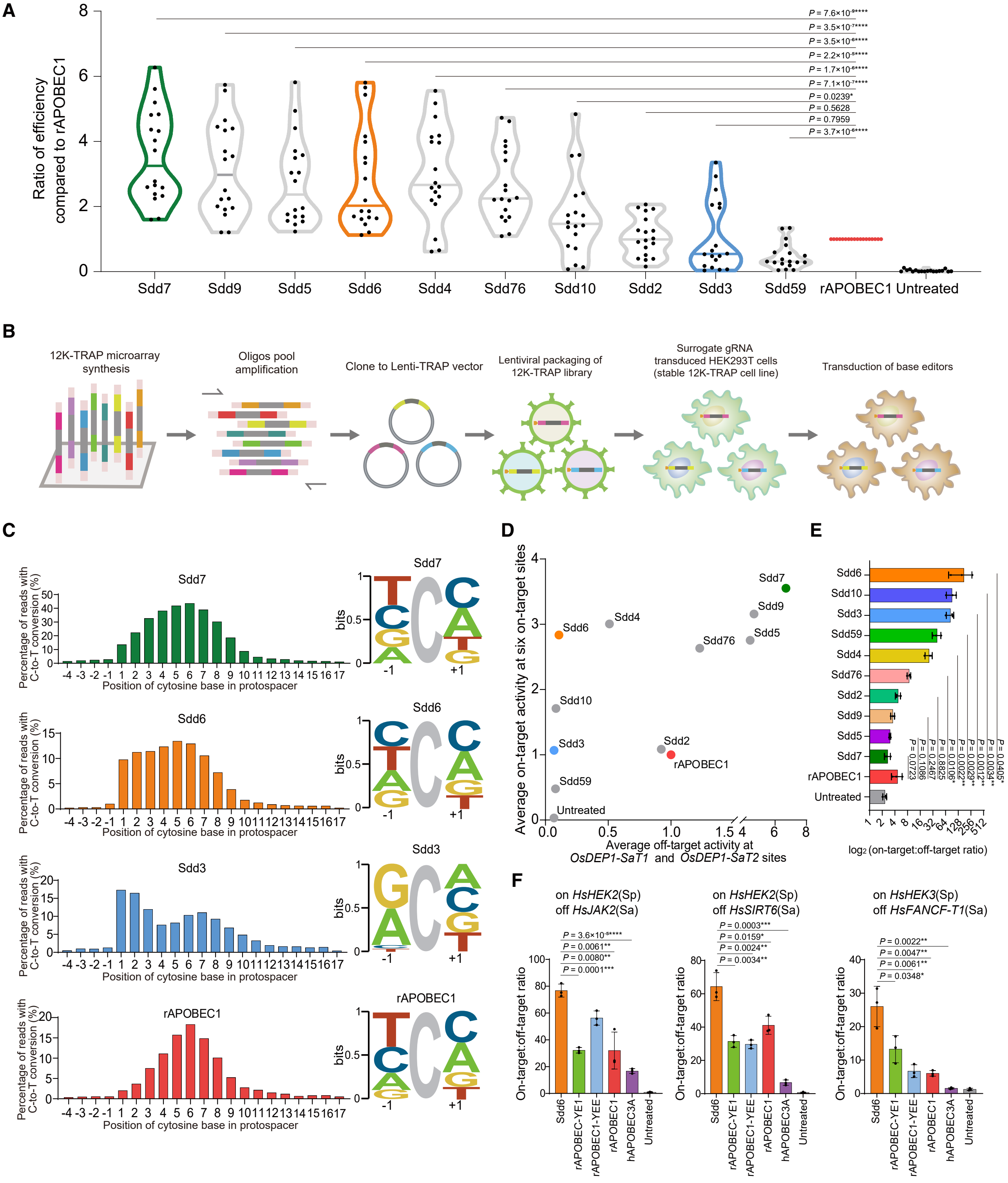


Figure 4. Evaluating newly discovered Sdd proteins for use as base editors in plant and human cells

(A) Overall editing efficiencies of the Sdds and rAPOBEC1 across six endogenous target sites in rice protoplasts. The average editing frequencies using rAPOBEC1 at each target were set to 1, and frequencies observed with Sdds were normalized accordingly. Dots represent each of three individual biological replicates across six endogenous genomic sites.

(B) Overview of using 12K-TRAP-seq to perform high-throughput quantification of the activities and properties of the Sdds and rAPOBEC1 in HEK293T cells.

(legend continued on next page)

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

deamination motif preferences, and altered editing windows, compared with conventional CBEss.

It was previously described that CBEs could cause genome-wide Cas9-independent off-target editing outcomes, which raises concerns about the safety of these precise genome editing technologies for clinical applications.[40,41] It is thought that these off-target mutations may be a result of overexpression of the cytidine deaminase. We wondered whether the newly discovered Sdd proteins could offer a more favorable balance between off-target and on-target editing. We therefore evaluated the Cas9-independent off-target effects of the 10 Sdds, using an established orthogonal R-loop assay in rice protoplasts.[42] We found that 6 (Sdd2, Sdd3, Sdd4, Sdd6, Sdd10, and Sdd59) of the 10 deaminases had lower off-target activities than rAPOBEC1. Interestingly, while Sdd6 showed nearly no off-target editing activity, it was still robust at on-target base editing when tested across six endogenous sites in rice protoplasts (Figures 4D and S4E). When we analyzed the on-target:off-target ratios of these 10 deaminases, Sdd6 exhibited the highest on-target:off-target editing ratio, which was 37.6-fold higher than that of rAPOBEC1 (Figure 4E). We further compared the on-target and off-target editing of Sdd6 to that of rAPOBEC1 and its two high-fidelity deaminase variants, YE1 and YEE, in HEK293T cells.[43] Importantly, we found that Sdd6 had the highest on-target:off-target editing ratios, which were calculated to be 2.8-, 2.1-, and 2.5-fold higher than that of rAPOBEC1, YE1, and YEE, respectively, and 10.4-fold higher than that of hA3A (Figures 4F and S4F). Notably, the on-target activity of Sdd6 was comparable to that of rAPOBEC1 and much higher than that of YE1 and YEE (Figure S4F). Thus, we identified that the SCP1.201 clade contains unique and more precise Sdd proteins to be used as high-fidelity base editors.

Rational truncation of Sdd proteins assisted by AlphaFold2 structure prediction

Although viral delivery of CBEs has great potential for disease treatment, the large size of APOBEC/AID-like deaminases restricts their ability to be packaged into single-AAV particles for *in vivo* editing applications.[31] Others have developed dual-AAV strategy delivery approaches by splitting CBEs into an amino-terminal and carboxy-terminal fragment and packaging them into separate AAV particles.[31] However, these delivery efforts would challenge large-scale manufacturing, require higher viral dosages, and would pose potential safety challenges for human use.[44] Recently, a truncated sea lamprey cytidine deaminase-like 1 (PmCDA1)-based CBE was developed that could theoretically be packaged into a single AAV, but the editing efficiency was extremely low when using the packaged AAVs for HEK293T cell transduction.[45] As SCP1.201 deaminases are canonically compact and conserved (Figure S5A), we thought that they might be the ideal protein for single-AAV CBEs.

We wondered whether we could use AI-assisted protein modeling to further engineer and shorten the size of the newly discovered Sdd proteins. We then generated multiple truncated variants of Sdd7, Sdd6, Sdd3, Sdd9, Sdd10, and Sdd4, and tested these variants for endogenous base editing in rice protoplasts across two sites each.

We identified mini-Sdd7, mini-Sdd6, mini-Sdd3, mini-Sdd9, mini-Sdd10, and mini-Sdd4 as newly minimized deaminases that are small (~130–160 aa) and have comparable or higher editing efficiencies, compared with their full-length proteins, both in rice protoplasts and human cells (Figures 5A and S5B–S5E). We found that the structures of these mini-proteins are conserved throughout the protein structural alignment (Figures 5A, S5D, and S5E). Strikingly, all six miniaturized deaminases would permit the construction of single-AAV-encapsulated SaCas9-based CBEs (<4.7 kb between inverted terminal repeats [ITRs]) (Figures 5B and S5F–S5H). We used mini-Sdd6 to construct a single-AAV SaCas9 vector and found that it had editing efficiencies of around 60% in mouse neuroblastoma N2a cells at two sites in the *Mus musculus* 4-hydroxyphenylpyruvate dioxygenase (*HPD*) gene[46] by transient transfection (Figure 5C). Following successful AAV packaging and tittering, we infected N2a cells and directed editing to the *MmHPD-T2* site. The editing efficiency significantly increased with increasing AAV concentrations, reaching up to 43.1% editing (Figure 5D). These results highlight that the Sdd proteins offer great advantages over APOBEC/AID-like deaminases in terms of AAV-based CRISPR base editing delivery. The success in further shortening Sdd proteins for AAV packaging highlights the great potential of AI-assisted protein engineering.

Robust base editing with Sdd-based CBEs in rice and soybean

We next explored the use and application of newly engineered Sdd proteins for base editing in plants. We first evaluated the use of mini-Sdd7 in *Agrobacterium*-mediated genome editing of rice plants and observed more mutants recovered and a greater proportion of edited plants, which reflects both a higher efficiency and lower toxicity, compared with the most used hA3A-based CBE in agricultural application (Figure S5I).

Soybean is one of the most important staple crops grown around the world, serving as an essential source of vegetable oil and protein.[47] Although previously reported base editors

(C) Overview of the editing properties and patterns of the Sdds and rAPOBEC1, as evaluated by the 12K-TRAP library. Left, the editing efficiencies and editing windows of the deaminases. Right, a sequence motif logo reflecting the context preferences of the deaminases.

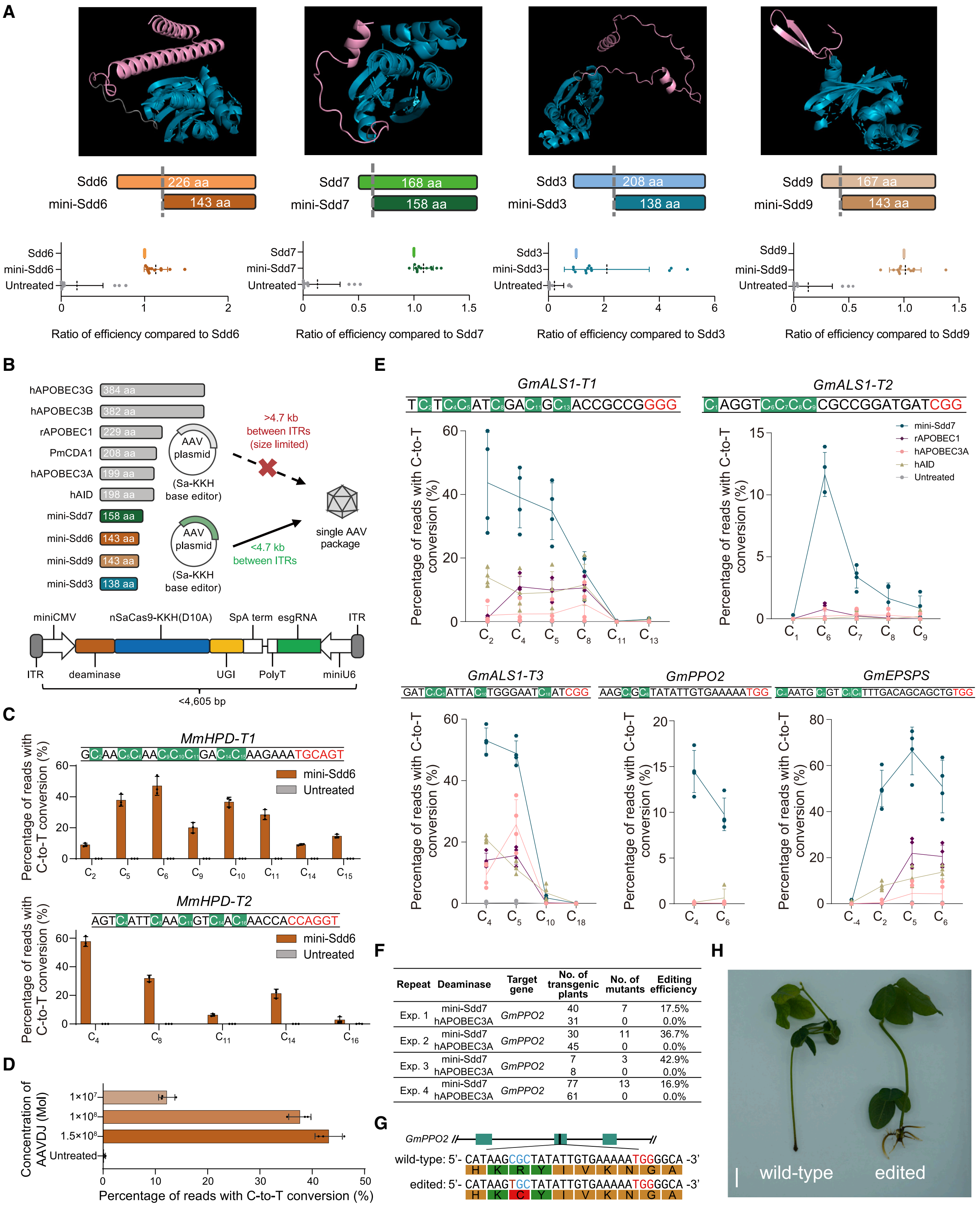
(D) Evaluation of off-target effects using an orthogonal R-loop assay in rice protoplasts. Dots represent average on-target C-to-T conversion frequencies of three independent biological replicates across six on-target sites in rice in (A) versus average sgRNA-independent off-target C-to-T conversion frequencies across two ssDNA regions (*OsDEP1-SaT1* and *OsDEP1-SaT2*) for each base editor.

(E) On-target:off-target editing ratios for each base editor calculated from (D).

(F) On-target:off-target editing ratios of Sdd6, rAPOBEC1-YE1, rAPOBEC1-YEE, rAPOBEC1, and hAPOBEC3A tested across two on-target and three off-target sites in HEK293T cells.

For (E) and (F), dots represent individual biological replicates, bars represent mean values, and error bars represent the SD of three independent biological replicates (n = 3). Data are presented as mean values ± SD. p values were obtained using two-sided Mann-Whitney tests. *p < 0.05, **p < 0.01, ***p < 0.001, and ****p < 0.0001.

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041



Repeat	Deaminase	Target gene	No. of transgenic plants	No. of mutants	Editing efficiency
Exp. 1	mini-Sdd7	GmPPO2	40	7	17.5%
	hAPOBEC3A		31	0	0.0%
Exp. 2	mini-Sdd7	GmPPO2	30	11	36.7%
	hAPOBEC3A		45	0	0.0%
Exp. 3	mini-Sdd7	GmPPO2	7	3	42.9%
	hAPOBEC3A		8	0	0.0%
Exp. 4	mini-Sdd7	GmPPO2	77	13	16.9%
	hAPOBEC3A		61	0	0.0%

(legend on next page)

have been widely used in many crops like rice, wheat, maize, potato, and more, cytosine base editing remains challenging and poorly efficient across most sites tested in soybean crops.[32,48] Since the first development of base editing, only one article has used *Agrobacterium tumefaciens* to obtain stable transformations and cytosine base-edited soybeans, but the efficiency was extremely low and resulted in chimeric plants rather than completely edited soybeans.[32]

We wondered whether our newly developed Sdd-based CBEs would result in superior cytosine base editing in soybeans. The transient base editing shown was evaluated using a soybean hairy root transformation mediated by *Agrobacterium rhizogenes*. This approach is often used in soybeans due to its quick nature (~20 days) in allowing researchers to evaluate editing percentages in root cells. We constructed vectors with an AtU6 promoter driving single-guide RNA molecule (sgRNA) expression and a Cauliflower Mosaic Virus (CaMV) 2 × 35S promoter driving CBE expression, and we evaluated these by using transgenic soybean hairy roots following *Agrobacterium rhizogenes*-mediated transformations (Figure S5J). We found that the APOBEC/AID-like deaminases had low editing activities across all five sites evaluated, as expected, including at the *GmALS1-T2* and *GmPPO2* sites that were particularly difficult to edit by other CBEs in soybean (Figure 5E). Remarkably, mini-Sdd7 displayed 26.3-, 28.2-, and 10.8-fold increased cytosine base editing levels, compared with rAPOBEC1, hA3A, and human AID (hAID), respectively, across the 5 sites and reached editing efficiencies up to 67.4% (Figure 5E). However, the cells from hairy root transformations are impossible to regenerate into soybean plants, so the canonical *Agrobacterium tumefaciens* is used to perform stable soybean plant editing in cotyledons.

We next sought to use hA3A and mini-Sdd7 to base edit and obtain transgenic soybean plants following *Agrobacterium tumefaciens*-mediated transformation. We chose to edit the endogenous *GmPPO2* gene to create an R98C mutation, which would result in carfentrazone-ethyl-resistant soybean plants.[49] Although the editing efficiencies from hairy root transformations are a great approach for evaluating relative editing efficiencies, they are not reflective of the percentage of edited plants following soybean plant regeneration. Even with the highly efficient hA3A base editor in plants, we never successfully obtained cytosine base-edited plants (Figure 5F). Surprisingly, we obtained 34 base-edited heterozygotes from 154 transgenic soybean seedlings of Sdd7 transgenic plants from 4 independent biological experiments (Figure 5F). Therefore, Sdd7 now enables efficient cytosine base editing in soybean plants, which will greatly contribute to future agricultural breeding efforts (Figures 5F and 5G).

After treatment with carfentrazone-ethyl for 10 days, we could obviously observe that while the wild-type plant was sensitive to wilting and could not generate roots, the mutated plant edited by Sdd7 grew well and normal (Figure 5H). The development of efficient CBEs for use in soybean plants could enable diverse applications in the future.

DISCUSSION

Compared with the limited insights provided by 1D amino acid sequence alone, 3D structural information provides a more visually informative representation of potential protein functions. Structure-based protein mining promises to be a useful method for discovering and engineering new enzymes. Previously, research in functional genomics has been limited by either the cost of high-resolution analysis of protein structure or by the low-accuracy of traditional computational-driven folding simulations.[50,51] AI-based high-accuracy protein folding prediction models and the related databases have breathed new life into the life sciences.

Here, we carried out a proof-of-concept exploration of protein classification and mining of protein functions, based on structural predictions for the cytidine deaminase-like superfamily. We showed that AlphaFold2-predicted structures classified deaminases reliably into distinct clades with diverse protein folds and catalytic functions. We built on this by identifying deaminases with novel and different DNA substrates, which in turn permits the design of bespoke precision genome editing tools. In principle, this strategy could be applied to the high-throughput classification and functional analysis of any protein dataset. We believe that future sequencing efforts in parallel with structural predictions will substantially advance the mining, tracking, classification, and design of functional proteins.

Currently only a few cytidine deaminases are in use as CBEs. Canonical efforts based solely on protein engineering and directed evolution have helped to diversify editing properties; however, these efforts are generally difficult to establish. Using

Figure 5. Engineering truncated Sdd proteins for use in animals and plants

(A) Engineering truncated Sdd proteins. Top, AlphaFold2-predicted structures of Sdd6, Sdd7, Sdd3, and Sdd9. Conserved regions are shown in cyan, and truncated regions are shown in pink. Bottom, relative editing efficiencies of Sdds and their minimized version across two endogenous sites in rice protoplasts and two sites in HEK293T cells.

(B) Theoretical packaging of a SaCas9-based CBE vector for packaging into a single AAV. Top, schematic diagram of APOBEC/AID-like deaminases, Sdds, and their AAV vectors. Gray-colored deaminases are too large for single-AAV packaging. Bottom, schematic representation of Sdd-based AAV vectors.

(C) Editing efficiency of mini-Sdd6 at two endogenous target sites in the *MmHPD* gene in N2a cells.

(D) AAV infection efficiency in N2a cell line at *MmHPD-T2* site. Three AAV concentrations were tested.

(E) Editing efficiencies of mini-Sdd7, rAPOBEC1, hA3A, and hAID base editors at five endogenous target sites in soybean hairy roots.

(F) Frequencies of mutations induced by mini-Sdd7 and hA3A in T_0 stable soybean plant editing in cotyledons by canonical *Agrobacterium tumefaciens*. The data were collected by four independent biological experiments.

(G) The genotypes of base-edited soybean plants.

(H) Phenotypes of soybean plants treated with carfentrazone-ethyl for 10 days. Left, wild-type soybean plant (R98). Right, base-edited soybean plant (C98). Scale bars, 1 cm.

For (A) and (C)–(E), dots represent individual biological replicates, bars represent mean values, and error bars represent the SD of three or four independent biological replicates.

our structure-based clustering methods, we discovered and profiled a suite of deaminases with distinct properties that can work both in plants and mammalian cells.

Among the AI-rational-discovered and -designed deaminases, we identified compacted Sdd7 and Sdd6 that show great promise for both therapeutic and agricultural applications. Sdd7 was capable of robust base editing in all tested species and had much higher editing activity than the most commonly used APOBEC/AID-like deaminases. Surprisingly, we found that Sdd7 was capable of efficiently editing soybean plants, which has been a major limitation for cytosine base editing previously. We speculated that Sdd7, derived from the bacterium *Actinosynnema mirum*, may possess high activity at temperatures suitable for soybean growth, in contrast to the mammalian APOBEC/AID-like deaminases. While profiling Sdd6, we found that this deaminase was smaller and by default more specific than the other deaminases, while maintaining high on-target editing activity. We believe that these newer discoveries and engineering efforts will contribute to the development of bespoke genome editing tools, which will be more precise and specific to each therapeutic or breeding application.

Advances in sequencing methods have propelled the discovery of new species and proteins. We also believe that AI-guided protein structure prediction and classification will provide a new effective perspective for protein classification with variable sequences and low sequence conservation, such as immune-related proteins. The advent of AI-assisted protein structure predictions in combination with growing numbers of sequencing efforts will further spark new enzyme discovery and enable even greater bioengineering efforts.

Limitations of the study

The classification method based on 3D structure alignment demonstrates great advantages but still has some limitations. Firstly, it is not suitable for proteins with high sequence identity. For functional differences caused by SNPs, or for proteins with high sequence identity, the structural differences are difficult to be fully characterized by AlphaFold2 or other structure prediction methods. Secondly, it is not suitable for proteins that rely on oligomeric or multiple complexes or proteins that have out-of-phase features when active *in vivo*. For proteins with these variable dynamic processes, a combination of molecular dynamics simulation is required. Finally, it is not suitable for proteins with high difficulty in obtaining precise structures based on predictive methods. For example, AlphaFold2 does not provide a high-confidence prediction result for many orphan proteins, and it is believed that with the development of new algorithms, this problem can be effectively solved.

Furthermore, due to the length and time constraints of this paper, we cannot fully explore the properties of all proteins in the SCP1.201 family and other family proteins. However, we believe that in future studies, there will be many surprises for these large and unknown protein families.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2023.05.041.

ACKNOWLEDGMENTS

We thank Prof. Youwei Ai and Prof. Qingfeng Wu for kindly providing HEK293T and N2a cell lines, respectively. We thank Prof. Qi-jun Chen for kindly providing pBSE901. We thank Prof. Tianfu Han for kindly providing the seeds of Zhonghuang13 soybean. This work was supported by the National Natural Science Foundation of China (32388201), the National Key Research and Development Program (2022YFF1002802), the Ministry of Agriculture and Rural Affairs of China, and the Strategic Priority Research Program of the Chinese Academy of Sciences (Precision Seed Design and Breeding, XDA24020102). Q.L. is supported by the Postdoctoral Innovative Talent Support Program of China (BX2021353) and China Postdoctoral Science Foundation (2022M720163). R.Z. is supported by the Youth Innovation Promotion Association CAS (2022097). K.T.Z. was supported by the Schmidt Science Fellows.

AUTHOR CONTRIBUTIONS

J. Huang, K.T.Z., and C.G. conceived the project and designed the experiments. J. Huang discovered the deaminases. H.F. and Y. Li performed the structure-based protein classification analysis. J. Huang, Q.L., and Z.H. performed the protoplasts transformation and NGS data collection experiments. J. Huang and Q.L. performed the mammalian cell transfection and NGS data collection experiments. J. Huang, H.F., and Z.H. collected the TRAP-seq data. J. Huang and Q.G. analyzed the Novaseq and Miseq data. G.L. and J. Hu prepared Miseq samples. H.F. and G.L. constructed the binary vectors for rice and soybean plant transformation. B.L. obtained regenerated rice plants. J. Huang, Z.H., and B.L. identified rice mutants. H.X., H.F., L.Z., Y.R.,

Z.H., and R.Z. performed soybean transformation, base-edited plants identification, and soybean resistance experiments. Y. Luo, K.Q., and P.H. generated the HEK293T cells with stable transfected 12K-TRAP-seq library. E.Z. provided AAV vector with guide RNA for mouse targets. Q.L. and Z.H. prepared the figures. J.-L.Q. revised the manuscript. C.G. and K.T.Z. supervised the study. Q.L., H.F., Y. Li, K.T.Z., and C.G. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors have submitted two patent applications based on the results reported in this paper. C.G. is a member of the *Cell* advisory board. K.T.Z. is a founder and employee at Qi Biodesign.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: March 7, 2023
Revised: April 24, 2023
Accepted: May 26, 2023
Published: June 27, 2023

REFERENCES

1. Sharifi, F., and Ye, Y. (2022). Identification and classification of reverse transcriptases in bacterial genomes and metagenomes. Nucleic Acids Res. *50*, e29. https://doi.org/10.1093/nar/gkab1207.
2. Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol. *18*, 67–83. https://doi.org/10.1038/s41579-019-0299-x.
3. Berntsson, R.P., Smits, S.H., Schmitt, L., Slotboom, D.J., and Poolman, B. (2010). A structural classification of substrate-binding proteins. FEBS Lett. *584*, 2606–2617. https://doi.org/10.1016/j.febslet.2010.04.043.
4. Chandonia, J.M., Guan, L., Lin, S., Yu, C., Fox, N.K., and Brenner, S.E. (2022). SCOPe: improvements to the structural classification of proteins—extended database to facilitate variant interpretation and machine learning. Nucleic Acids Res. 50, D553–D559.
5. wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. *47*, D520–D528. https://doi.org/10.1093/nar/gky949.
6. Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M.L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L.J., et al. (2023). MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res. *47*, D520–D528. https://doi.org/10.1093/nar/gkac1080.
7. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
8. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.
9. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.
10. Mok, B.Y., de Moraes, M.H., Zeng, J., Bosch, D.E., Kotrys, A.V., Raguram, A., Hsu, F., Radey, M.C., Peterson, S.B., Mootha, V.K., et al. (2020). A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. Nature *583*, 631–637. https://doi.org/10.1038/s41586-020-2477-4.
11. Zhang, H., Yang, B., Pomerantz, R.J., Zhang, C., Arunachalam, S.C., and Gao, L. (2003). The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature *424*, 94–98. https://doi.org/10.1038/nature01707.
12. Weiss, B. (2007). The deoxycytidine pathway for thymidylate synthesis in *Escherichia coli*. J. Bacteriol. *189*, 7922–7926. https://doi.org/10.1128/JB.00461-07.
13. Esnault, C., Heidmann, O., Delebecque, F., Dewannieux, M., Ribet, D., Hance, A.J., Heidmann, T., and Schwartz, O. (2005). APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. Nature *433*, 430–433. https://doi.org/10.1038/nature03238.
14. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424. https://doi.org/10.1038/nature17946.
15. Nishida, K., Arazoe, T., Yachie, N., Banno, S., Kakimoto, M., Tabata, M., Mochizuki, M., Miyabe, A., Araki, M., Hara, K.Y., et al. (2016). Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science *353*, aaf8729. https://doi.org/10.1126/science.aaf8729.
16. Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J., and Zhang, F. (2017). RNA editing with CRISPR-Cas13. Science *358*, 1019–1027. https://doi.org/10.1126/science.aaq0180.
17. Harris, R.S., Petersen-Mahrt, S.K., and Neuberger, M.S. (2002). RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. Mol. Cell *10*, 1247–1253. https://doi.org/10.1016/s1097-2765(02)00742-6.
18. Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. Nature *550*, 249–254. https://doi.org/10.1038/nature24041.
19. Wolf, J., Gerber, A.P., and Keller, W. (2002). tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. EMBO J. *21*, 3841–3851. https://doi.org/10.1093/emboj/cdf362.
20. Iyer, L.M., Zhang, D., Rogozin, I.B., and Aravind, L. (2011). Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. Nucleic Acids Res. *39*, 9473–9497. https://doi.org/10.1093/nar/gkr691.
21. Krishnan, A., Iyer, L.M., Holland, S.J., Boehm, T., and Aravind, L. (2018). Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity. Proc. Natl. Acad. Sci. USA *115*, E3201–E3210. https://doi.org/10.1073/pnas.1720897115.
22. Gao, C. (2021). Genome engineering for crop improvement and future agriculture. Cell *184*, 1621–1635. https://doi.org/10.1016/j.cell.2021.01.005.
23. Anzalone, A.V., Koblan, L.W., and Liu, D.R. (2020). Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. Nat. Biotechnol. *38*, 824–844. https://doi.org/10.1038/s41587-020-0561-9.
24. Li, Y., Li, W., and Li, J. (2021). The CRISPR/Cas9 revolution continues: from base editing to prime editing in plant science. J. Genet. Genomics *48*, 661–670. https://doi.org/10.1016/j.jgg.2021.05.001.
25. Zhang, R., Chen, S., Meng, X., Chai, Z., Wang, D., Yuan, Y., Chen, K., Jiang, L., Li, J., and Gao, C. (2021). Generating broad-spectrum tolerance to ALS-inhibiting herbicides in rice by base editing. Sci. China Life Sci. *64*, 1624–1633. https://doi.org/10.1007/s11427-020-1800-5.
26. Chen, Y., Wang, Z., Ni, H., Xu, Y., Chen, Q., and Jiang, L. (2017). CRISPR/Cas9-mediated base-editing system efficiently generates gain-of-function mutations in Arabidopsis. Sci. China Life Sci. *60*, 520–523. https://doi.org/10.1007/s11427-017-9021-5.

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

27. Ma, Y., Zhang, J., Yin, W., Zhang, Z., Song, Y., and Chang, X. (2016). Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat. Methods *13*, 1029–1035. https://doi.org/10.1038/nmeth.4027.

28. Hess, G.T., Frésard, L., Han, K., Lee, C.H., Li, A., Cimprich, K.A., Montgomery, S.B., and Bassik, M.C. (2016). Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat. Methods *13*, 1036–1042. https://doi.org/10.1038/nmeth.4038.

29. Yu, Y., Leete, T.C., Born, D.A., Young, L., Barrera, L.A., Lee, S.J., Rees, H.A., Ciaramella, G., and Gaudelli, N.M. (2020). Cytosine base editors with minimized unguided DNA and RNA off-target events and high on-target activity. Nat. Commun. *11*, 2052. https://doi.org/10.1038/s41467-020-15887-5.

30. Cheng, T.L., Li, S., Yuan, B., Wang, X., Zhou, W., and Qiu, Z. (2019). Expanding C-T base editing toolkit with diversified cytidine deaminases. Nat. Commun. *10*, 3612. https://doi.org/10.1038/s41467-019-11562-6.

31. Levy, J.M., Yeh, W.H., Pendse, N., Davis, J.R., Hennessey, E., Butcher, R., Koblan, L.W., Comander, J., Liu, Q., and Liu, D.R. (2020). Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. Nat. Biomed. Eng. *4*, 97–110. https://doi.org/10.1038/s41551-019-0501-5.

32. Cai, Y., Chen, L., Zhang, Y., Yuan, S., Su, Q., Sun, S., Wu, C., Yao, W., Han, T., and Hou, W. (2020). Target base editing in soybean using a modified CRISPR/Cas9 system. Plant Biotechnol. J. *18*, 1996–1998. https://doi.org/10.1111/pbi.13386.

33. Sokal, R.R., and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. Kansas Univ. Sci. Bull. *38*, 1409–1438.

34. Zong, Y., Wang, Y., Li, C., Zhang, R., Chen, K., Ran, Y., Qiu, J.L., Wang, D., and Gao, C. (2017). Precise base editing in rice, wheat and maize with a Cas9-cytidine deaminase fusion. Nat. Biotechnol. *35*, 438–440. https://doi.org/10.1038/nbt.3811.

35. Mok, B.Y., Kotrys, A.V., Raguram, A., Huang, T.P., Mootha, V.K., and Liu, D.R. (2022). CRISPR-free base editors with enhanced activity and expanded targeting scope in mitochondrial and nuclear DNA. Nat. Biotechnol. *40*, 1378–1387. https://doi.org/10.1038/s41587-022-01256-8.

36. Koblan, L.W., Doman, J.L., Wilson, C., Levy, J.M., Tay, T., Newby, G.A., Maianti, J.P., Raguram, A., and Liu, D.R. (2018). Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. Nat. Biotechnol. *36*, 843–846. https://doi.org/10.1038/nbt.4172.

37. Zong, Y., Song, Q., Li, C., Jin, S., Zhang, D., Wang, Y., Qiu, J.L., and Gao, C. (2018). Efficient C-to-T base editing in plants using a fusion of nCas9 and human APOBEC3A. Nat. Biotechnol. *36*, 950–953. https://doi.org/10.1038/nbt.4261.

38. Lin, Q., Zhu, Z., Liu, G., Sun, C., Lin, D., Xue, C., Li, S., Zhang, D., Gao, C., Wang, Y., et al. (2021). Genome editing in plants with MAD7 nuclease. J. Genet. Genomics *48*, 444–451. https://doi.org/10.1016/j.jgg.2021.04.003.

39. Xiang, X., Qu, K., Liang, X., Pan, X., Wang, J., Han, P., Dong, Z., Liu, L., Zhong, J., Ma, T., Wang, Y., et al. (2020). Massively parallel quantification of CRISPR editing in cells by TRAP-seq enables better design of Cas9, ABE, CBE gRNAs of high efficiency and accuracy https://doi.org/10.1101/2020.05.20.103614.

40. Jin, S., Zong, Y., Gao, Q., Zhu, Z., Wang, Y., Qin, P., Liang, C., Wang, D., Qiu, J.L., Zhang, F., et al. (2019). Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. Science *364*, 292–295. https://doi.org/10.1126/science.aaw7166.

41. Zuo, E., Sun, Y., Wei, W., Yuan, T., Ying, W., Sun, H., Yuan, L., Steinmetz, L.M., Li, Y., and Yang, H. (2019). Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. Science *364*, 289–292. https://doi.org/10.1126/science.aav9973.

42. Jin, S., Fei, H., Zhu, Z., Luo, Y., Liu, J., Gao, S., Zhang, F., Chen, Y.H., Wang, Y., and Gao, C. (2020). Rationally designed APOBEC3B cytosine base editors with improved specificity. Mol. Cell *79*, 728–740.e6. https://doi.org/10.1016/j.molcel.2020.07.005.

43. Doman, J.L., Raguram, A., Newby, G.A., and Liu, D.R. (2020). Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. Nat. Biotechnol. *38*, 620–628. https://doi.org/10.1038/s41587-020-0414-6.

44. Davis, J.R., Wang, X., Witte, I.P., Huang, T.P., Levy, J.M., Raguram, A., Banskota, S., Seidah, N.G., Musunuru, K., and Liu, D.R. (2022). Efficient in vivo base editing via single adeno-associated viruses with size-optimized genomes encoding compact adenine base editors. Nat. Biomed. Eng. *6*, 1272–1283. https://doi.org/10.1038/s41551-022-00911-4.

45. Li, A., Mitsunobu, H., Yoshioka, S., Suzuki, T., Kondo, A., and Nishida, K. (2022). Cytosine base editing systems with minimized off-target effect and molecular size. Nat. Commun. *13*, 4531. https://doi.org/10.1038/s41467-022-32157-8.

46. Pankowicz, F.P., Barzi, M., Legras, X., Hubert, L., Mi, T., Tomolonis, J.A., Ravishankar, M., Sun, Q., Yang, D., Borowiak, M., et al. (2016). Reprogramming metabolic pathways in vivo with CRISPR/Cas9 genome editing to treat hereditary tyrosinaemia. Nat. Commun. *7*, 12642. https://doi.org/10.1038/ncomms12642.

47. Liu, S., Zhang, M., Feng, F., and Tian, Z. (2020). Toward a "Green Revolution" for soybean. Mol. Plant *13*, 688–697. https://doi.org/10.1016/j.molp.2020.03.002.

48. Molla, K.A., Sretenovic, S., Bansal, K.C., and Qi, Y. (2021). Precise plant genome editing using base editors and prime editors. Nat. Plants *7*, 1166–1187. https://doi.org/10.1038/s41477-021-00991-1.

49. Dayan, F.E., Barker, A., and Tranel, P.J. (2018). Origins and structure of chloroplastic and mitochondrial plant protoporphyrinogen oxidases: implications for the evolution of herbicide resistance. Pest Manag. Sci. *74*, 2226–2234. https://doi.org/10.1002/ps.4744.

50. Thompson, M.C., Yeates, T.O., and Rodriguez, J.A. (2020). Advances in methods for atomic resolution macromolecular structure determination. F1000Res *9*, 667. https://doi.org/10.12688/f1000research.25097.1.

51. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710. https://doi.org/10.1038/s41586-019-1923-7.

52. Shan, Q., Wang, Y., Li, J., and Gao, C. (2014). Genome editing in rice and wheat using the CRISPR/Cas system. Nat. Protoc. *9*, 2395–2410. https://doi.org/10.1038/nprot.2014.157.

53. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. (2009). InterPro: the integrative protein signature database. Nucleic Acids Res. *37*, D211–D215. https://doi.org/10.1093/nar/gkn785.

54. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. *36*, W5–W9. https://doi.org/10.1093/nar/gkn201.

55. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37. https://doi.org/10.1093/nar/gkr367.

56. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

57. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

58. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

59. Zhang, C., Shine, M., Pyle, A.M., and Zhang, Y. (2022). US-align: universal structure alignments of proteins, nucleic acids, and macromolecular

complexes. Nat. Methods *19*, 1109–1115. https://doi.org/10.1038/s41592-022-01585-1.

60. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. A *34*, 827–828. https://doi.org/10.1107/S0567739478001680.

61. DeLano, W.L. (2000). The PyMOL Molecular Graphics System (Schrödinger LLC).

62. Jin, S., Lin, Q., Gao, Q., and Gao, C. (2023). Optimized prime editing in monocot plants using PlantPegDesigner and engineered plant prime editors (ePPEs). Nat. Protoc. *18*, 831–853. https://doi.org/10.1038/s41596-022-00773-9.

63. Jin, S., Gao, Q., and Gao, C. (2021). An unbiased method for evaluating the genome-wide specificity of base editors in rice. Nat. Protoc. *16*, 431–457. https://doi.org/10.1038/s41596-020-00423-y.

64. Li, C., Zhang, H., Wang, X., and Liao, H. (2014). A comparison study of *Agrobacterium*-mediated transformation methods for root-specific promoter analysis in soybean. Plant Cell Rep. *33*, 1921–1932. https://doi.org/10.1007/s00299-014-1669-5.

65. Li, S., Cong, Y., Liu, Y., Wang, T., Shuai, Q., Chen, N., Gai, J., and Li, Y. (2017). Optimization of Agrobacterium-mediated transformation in soybean. Front. Plant Sci. *8*, 246. https://doi.org/10.3389/fpls.2017.00246.

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
FastT1 Competent Cells	Vazyme	C505-02
Chemicals, peptides, and recombinant proteins		
DMEM (1X)+GlutaMax	Gibco	10569044
FBS Fetal Bovine Serum, Qualified	Gibco	10091148
TrypLE Express	Gibco	12605-010
PBS pH 7.4 basic (1X)	Gibco	C10010500BT
Streptomycin, Penicillin	Gibco	15140-122
Blasticidin S HCl	Gibco	A1113903
Puromycin	Gibco	A1113803
Opti-MEM	Gibco	31985-070
Lipofectamine 2000	Invitrogen	11668019
Trypan Blue stain 0.4%	Invitrogen	T10282
Countess cell counting chamber slides	Invitrogen	C10283
Phanta Max Master Mix	Vazyme	P525-01
Poly(ethylene glycol)	Sigma-Aldrich	81240
Magnesium chloride hexahydrate	Sigma-Aldrich	M9272
Calcium chloride dihydrate	Sigma-Aldrich	C7902
Potassium chloride	Sigma-Aldrich	P3911
Sodium chloride	Sigma-Aldrich	S5886
YEAST EXTRACT	OXOID	LP0021B
TRYPTONE	OXOID	LP0042B
Critical Commercial Assays		
Plasmid Plus Midi Kit (100)	QIAGEN	12945
Mycoplasma Detection Kit	Transgen	FM311-01
Cell/Tissue DNA Isolation Mini Kit	Vazyme	DC102-01
Triumfi Mouse Tissue Direct PCR Kit	Genesand	SD312
GeneJET Gel Extraction Kit	Thermo Scientific	K0692
Plant Genomic DNA Kit	Tiangen	DP305
PureYield™ Plasmid Miniprep System	Promega	A1222
Deposited data		
Deep amplicon sequencing data of rice	This paper	SRA: PRJNA915939
Deep amplicon sequencing data of human	This paper	SRA: PRJNA915940
Deep amplicon sequencing data of mouse	This paper	SRA: PRJNA915941
Deep amplicon sequencing data of soybean	This paper	SRA: PRJNA915942
Code for deep amplicon sequencing data analyses from NovaSeq platform	This paper	https://github.com/ReiGao/GEanalysis/tree/master/Scripts
Code for deep amplicon sequencing data analyses from Miseq platform	This paper	https://github.com/ReiGao/Miseq_BEanalysis
Experimental models: Cell lines		
HEK293T	ATCC	CRL-3216
N2a	ATCC	CCL-131
Oligonucleotides		
See Table S4	This paper	N/A

(Continued on next page)

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Recombinant DNA		
Ubi-BFP(TC)	This paper	N/A
Ubi-BFP(CC)	This paper	N/A
Ubi-BFP(AC)	This paper	N/A
Ubi-BFP(GC)	This paper	N/A
pnCas9-PBE	Zong et al.[34]	Addgene # 98164
pnCas9-deaminase-PBE	This paper	N/A
pOsU3	Lin et al.[38]	Addgene # 170132
pCMV_BE4max	Koblan et al.[36]	Addgene # 112093
pCMV-deaminase-BE4max	This paper	N/A
phU6	This paper	N/A
pCMV-NLS-TALE-deaminase-UGI-NLS	This paper	N/A
pJAK2target-NCN	This paper	N/A
A3A-PBE	Zong et al.[37]	Addgene # 119768
pCMV-A3A-BE4max	This paper	N/A
12K TRAP-seq library plasmid	Xiang et al.[39]	N/A
Ubi-nSaCas9	Jin et al.[42]	N/A
pOsU3-Sa	This paper	N/A
pCMV-nSaCas9	This paper	N/A
phU6-Sa	This paper	N/A
pCMV-YE1-BE4max	This paper	N/A
pCMV-YEE-BE4max	This paper	N/A
pCMV-mini-Sdd6-SaBE4max-bGH	This paper	N/A
pEFS-mini-Sdd6-SaBE4max-bGH	This paper	N/A
pminiCMV-mini-Sdd6-SaBE4max-bGH	This paper	N/A
pU1a-mini-Sdd6-SaBE4max-bGH	This paper	N/A
pCMV-mini-Sdd6-SaBE4max-spA	This paper	N/A
phminiU6-Sa	This paper	N/A
pH-Ubi-deam-nCas9-PBE	This paper	N/A
pE-35S-deam-nCas9-PBE	This paper	N/A
Software and algorithms		
GraphPad Prism 8	GraphPad Prism software	N/A
Adobe Illustrator	Adobe	N/A
PyMOL	DeLano Scientific LLC	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Caixia Gao (cxgao@genetics.ac.cn).

Materials availability

All unique/stable reagents generated in this study are available from the lead contact with a completed Materials Transfer Agreement.

Data and code availability

All deep amplicon sequencing data has been deposited at NCBI Sequence Read Archive database and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. Raw image data (Base editor reporter images) will be provided upon request from the lead contact.All original code has been deposited at GitHub and is publicly available. GitHub links are listed in the key resources table.

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

E.coli transfection

FastT1 *E.coli* competent cells were used for amplifying plasmid DNA. Transfected *E.coli* cells were grown at 37°C in Lysogeny Broth (LB) medium supplemented with 100 mg/mL ampicillin or kanamycin overnight.

Rice protoplast transfection

For protoplasts transfection, we used the Japonica rice (*Oryza sativa*) variety Zhonghua11 to prepare protoplasts. Protoplast isolation and transformation were performed as described previously.[52] Plasmids (5 μg per construct) were introduced by PEG-mediated transfection. The transfected protoplasts were normally incubated at 26 °C for 72 hours for fluorescence cell observation or DNA extraction.

Mammalian Cell lines and culture conditions

Both human HEK293T cells (ATCC, CRL-3216) and mouse N2a cells (ATCC, CCL-131) were cultured in Dulbecco's Modified Eagle's medium (DMEM, Gibco) supplemented with 10% (vol/vol) fetal bovine serum (FBS, Gibco) and 1% (vol/vol) Penicillin-Streptomycin (Gibco) in a humidified incubator at 37 °C with 5% CO_2. Cells were plated on 75 cm^2 Cell Culture Flask (NEST).

METHOD DETAILS

Protein clustering and analyzing

Protein sequences were downloaded from InterPro database[53] and NCBI's BLAST[54] (https://blast.ncbi.nlm.nih.gov/Blast.cgi) on the NR database. HMM was utilized to annotate deaminase domains to reduce the accumulation of unrelated information by HMMER.[55] We randomly chose 15 proteins from each family and clustered their domain sequences with a threshold of 90% sequence identity and 90% coverage using CD-HIT to reduce redundant sequences.[56] Representatives of each cluster were selected for further analysis. High confidence protein structures were predicted by Alphafold v2.2.0 and filtered with average per-residue confidence metric predicted local-distance difference test (pLDDT) $\geq$ 70.

Multiple sequence alignment was performed using Multiple Protein Sequence Alignment (MUSCLE).[57] The phylogenetic tree was constructed using IQ-TREE 2 (http://www.iqtree.org) with 1500 ultrafast bootstraps.[58] A low perturbation strength (-pers 0.2) and large number of stop iterations (-nstop 500) were set because of the short length of the deaminase domains. The paired structure alignment was performed based on TM-score method[59] and was further correction by relative distance between structural residues.[60] The overall structural similarity matrix was performed based on the results of the paired structure alignment and was further normalized sing the min-max method. The structural similarity matrix was further clustered by Unweighted Pair Group Method with Arithmetic mean (UPGMA)[33] and visualized by Figtree (http://tree.bio.ed.ac.uk/software/figtree/). Protein structure diagrams were made in PyMOL.[61]

Deaminase synthesis and removal of redundant sequence

We chose gene fragments encoding complete deaminase domains as well as extra N and C protein sequences for commercial synthesis (GenScript) (Figure S1A). All of the candidate cytidine deaminases were codon optimized (rice and wheat or human and mouse). The toxin deaminase was split into two fragments and the split site was selected according to DddA by protein structure alignment. The conserved protein structure was obtained through multiple alignment of predicted structure in PyMOL,[61] which helps to conduct the removal of redundant sequence.

Plasmid construction

For plant CBE vectors (maize ubiquitin-1 promoter-driven CBEs), synthesized deaminases were cloned into pnCas9-PBE vector (Addgene#98164), yielding vectors with Ubi-1::NLS-deaminase-linker-nCas9(D10A)-UGI-NLS::CaMV expression cassettes. Among them, pnCas9-miniSdd6-PBE, pnCas9-miniSdd7-PBE, and pnCas9-miniSdd3-PBE vectors are available in Addgene.

For CBE vectors for mammalian cells (CMV promoter-driven CBEs), synthesized deaminases-SpCas9-2UGI were cloned into p2T-CMV-ABEmax-BlastR vector (Addgene#152989), yielding vectors with CMV::NLS-deaminase-linker-nCas9(D10A)-2xUGI-NLS::bGH expression cassettes. Among them, p2T-CMV-miniSdd6-BE4max-BlastR, p2T-CMV-miniSdd7-BE4max-BlastR, and p2T-CMV-miniSdd3-BE4max-BlastR are available in Addgene.

The DdCBE vectors including NLS, TALE array sequences, candidate cytidine deaminases, and UGI sequence were codon optimized for both human and mouse, synthesized commercially (Genscript), and cloned into pCMV_BE4max vector (Addgene#112093), yielding vectors with CMV::NLS-TALE-deaminase-UGI-NLS::bGH expression cassettes. Among them, pCMV_TALE-L-JAK2-Ddd9-N and pCMV_TALE-R-JAK2-Ddd9-C vectors are available in Addgene.

The plant sgRNA vectors (rice U3 promoter drives sgRNA) were constructed as reported previously using the pOsU3 backbone (Addgene#170132).[62] To construct human and mouse sgRNA vectors (human U6 promoter drives sgRNA), the hU6 promoter was amplified and cloned into the pOsU3 backbone, followed by sgRNA target sequence cloning steps.[52]

Plant SaCas9 vectors for off-target testing were constructed as reported previously.[42]

Please cite this article in press as: Huang et al., Discovery of deaminase functions by structure-based protein clustering, Cell (2023), https://doi.org/10.1016/j.cell.2023.05.041

To construct AAV vectors, the sequences between ITRs were synthesized (GenScript) and cloned into pX601 vector (Addgene#61591), followed by sgRNA target sequence cloning steps. The AAV vector with *MmHPD-T2* sgRNA target sequence was named pAAV-EFS-SaKKH-spA-miniU6-miniSdd6-MmHPD-T2, this vector is available at Addgene.

To construct binary vectors for rice plant transformation, the candidate cytidine deaminases were codon optimized, synthesized commercially (GenScript), and cloned into pH-nCas9-PBE vector (Addgene#98163), followed by sgRNA target sequence cloning steps.[52] The vector without sgRNA target sequence was named pH-Ubi-miniSdd7-nCas9-PBE, this vector is available at Addgene.

To construct binary vectors for soybean hairy root transformation, NLS, candidate cytidine deaminases, linker, nCas9(D10A), UGI, P2A, mScarlet sequences were codon optimized, synthesized commercially (GenScript), and cloned into pBSE901 (Addgene#91709) vector, followed by sgRNA target sequence cloning steps. To construct binary vectors for soybean transformation, the selection marker was replaced by the *EPSPS* sequence. The vector with *GmPPO2* sgRNA target sequence and EPSPS selection marker was named pE-35S-miniSdd7-nCas9-PBE-GmPPO2, this vector is available at Addgene.

Mammalian cell line transfection

All the cells were routinely tested for Mycoplasma contamination with a Mycoplasma Detection Kit (Transgen Biotech). The cells were seeded into 48-well Clear TC-treated Plates plates (Corning) in the absence of antibiotic. After 16-24 hours, cells were incubated with 1 μL Lipofectamine 2000 (ThermoFisher Scientific), 300 ng vector with deaminases, and 100 ng sgRNA expression vector. For DdCBEs transfection, cells were incubated with 1 μL Lipofectamine 2000, 300ng TALE-L and 300ng TALE-R. 72 hours later the cells were washed with PBS, followed by DNA extraction. For examining off-target effects by the R-loop assay, four vectors namely BE4-max vector, SaCas9BE4max vector and the corresponding sgRNA vectors were co-transfected into cells.[36]

TRAPseq library

We used the sgRNA 12K-TRAPseq library for evaluation of base editor properties. We seeded 2×10^6 cells into 100 mm cell-culture dish (Corning) 20-hours before viral transduction. We transduced 500 μL of sgRNA lentivirus. For stably integrated cells, we used 1 μg/mL of puromycin (Gibco) to select. For each base editor, we seeded 2×10^6 cells into 6-plates 24-hours before transfection. We transfected 15 μg of each CBE member plasmid DNA and 15 μg of Tol2 DNA with 60 μL of Lipofectamine 2000. Following 24 hours after transfection, we changed new culturing media to contain 10 μg/mL blasticidin (Gibco). After another 3 days, we washed the cells, suspended and reseeded all cells in 10 μg/mL blasticidin-containing media. After 6 days, we harvested all cells by washing with PBS then centrifuged and extracted DNA using Cell/Tissue DNA Isolation Mini Kit (Vazyme). For each member, we prepared sequencing reactions by applying 1.2 μg of DNA with a first set of primers following by barcoding and next-generating sequencing.

DNA extraction

For HEK293T cells and N2a cells, genomic DNA was extracted with Lysis Buffer and Proteinase K with a Triumfi Mouse Tissue Direct PCR Kit (Beijing Genesand Biotech). For protoplasts, genomic DNA was extracted with a Plant Genomic DNA Kit (Tiangen Biotech) after 72 hours' incubation. All DNA samples were quantified with a NanoDrop 2000 spectrophotometer (Thermo Scientific).

Amplicon deep sequencing and data analysis

Triumfi Mouse Tissue Direct PCR Kit (Beijing Genesand Biotech) was used for amplification of target sequence in HEK293T cells and N2a cells. Phanta Max Master Mix (Vazyme) was used for amplification of target sequence in plants.

Nested PCR was used for amplification. In the first round PCR, the target region was amplified from genomic DNA with site-specific primers. In the second round, both forward and reverse barcodes were added to the ends of the PCR products for library construction. Equal amounts of PCR product were pooled and purified with a GeneJET Gel Extraction Kit (Thermo Scientific) and quantified with a NanoDrop 2000 spectrophotometer (Thermo Scientific). The purified products were sequenced commercially using the NovaSeq or Miseq platform, and the sequences around the target regions were examined for editing events.[62] The analysis pipeline of the sequencing data from NovaSeq platform can be refered as previous report[62] that has now been shared in GitHub website (https://github.com/ReiGao/GEanalysis/tree/master/Scripts) and the code of data processing from Miseq platform has been shared in GitHub website (https://github.com/ReiGao/Miseq_BEanalysis). Amplicon sequencing was repeated three times for each target site using genomic DNA extracted from three independent samples. Analysis of base editing behaviour by NovaSeq and Miseq was performed as described previously.[62]

For TRAP-seq analysis, we filtered next-generation sequencing (NGS) read depths of 12K TRAP below 50 and calculated the average editing efficiency at the corresponding surrogate target site inside the windows (from -10 to +27). In addition, we calculated the editing frequency for each NCN sequence motif and its proportions to evaluate context preferences.

All the primers used are listed in supplementary table (Table S5).

Agrobacterium-mediated transformation of rice calli

The Japonica rice (*Oryza sativa*) variety Zhonghua 11 was used for genetic transformation in this study. Binary vectors were introduced into *Agrobacterium tumefaciens* strain AGL1 by electroporation. *Agrobacterium*-mediated transformation of Zhonghua11 callus cells was conducted as reported.[63] Hygromycin (50 μg/ml) was used to select transgenic plants.

Soybean hairy root transformation and plant transformation

The soybean (*Glycine max*) variety Williams 82 was used to generate hairy roots. Binary vectors were introduced into *Agrobacterium rhizogenes* strain K599 by electroporation. Explants were allowed to grow and develop roots for around 20 days in germination medium. Transgenic hairy roots were generated without selection in 10-12 days.[64] The soybean (*Glycine max*) variety Zhonghuang13 were used for generation of transgenic plants using *Agrobacterium tumefaciens*-mediated stable transformation. 10 mg/L glyphosate was used for selection during plant regeneration.[65] For phenotype identification of base-edited soybean, 0.3 mg/L carfentrazone-ethyl were added in rooting medium for selection.

Plant mutant identification

Genomic DNA of transgenic plants was extracted with DNA Quick Plant System (Tiangen Biotech). Specific primers were used to amplify and sequence the target sites as described previously[62] (Table S5) (BGI). T_0 transgenic rice and soybean plants were examined individually.

Recombinant adeno-associated virus (rAAV) production and infection

The serotype type of AAV was selected as AAV-DJ, and the packaging and quantification was conducted at PackGene Biotech Company and the final viral titer is 1E+13 GC/mL. For infection, N2a cells seeded into 48-well Poly-D-Lysine-coated plates (Corning) of 50,000 cells/well with 250 μl of DMEM 24 hours before. To demonstrate the correlation between infection efficiency and concentration, we tested three concentrations, MOI: 1.0×10^7, 1.0×10^8 and 1.5×10^8. Cells were harvested 7 days after infection and the medium was changed on the third day. The genomic DNA was extracted using Triumfi Mouse Tissue Direct PCR Kit (Beijing Genesand Biotech).

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification

Datasets were assessed using GraphPad Prism 8 (GraphPad Software).

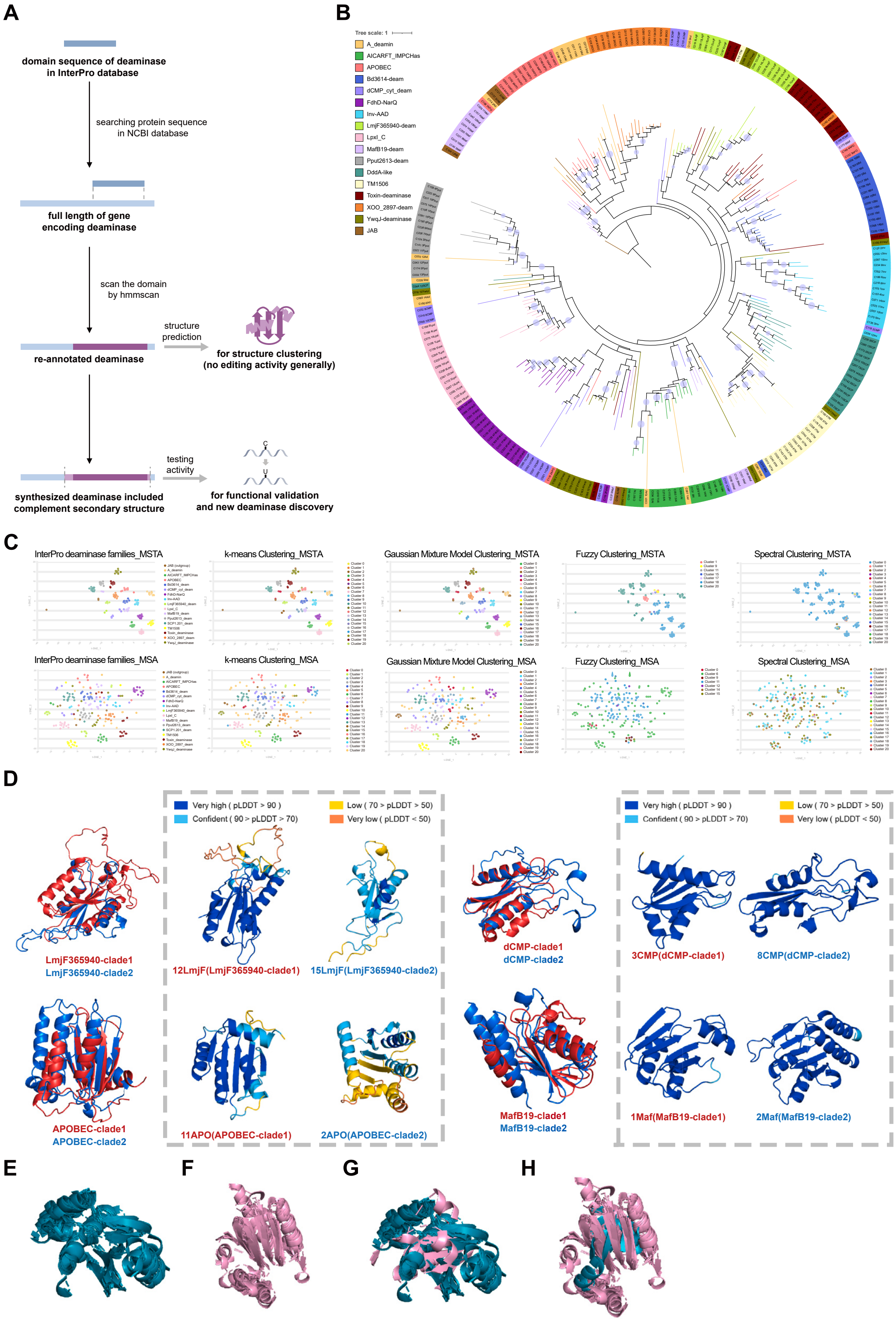
Statistical analysis

All numerical values are presented as means ± s.d. Differences between control and treatments were tested using Student's t-tests, and $P < 0.05$ was considered statistically significant, $P < 0.01$ was considered statistically extremely significant.

ADDITIONAL RESOURCES

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Supplemental figures



(legend on next page)

Figure S1. Discovery of cytidine deaminases via protein structures, related to Figure 1

(A) Workflow of re-annotation and synthesis of candidate deaminases. Since the amino acid sequence of the deaminase domains from InterPro may be incomplete, we used protein BLAST from the NCBI database to obtain the full length of gene encoding deaminase, and then re-annotated the deaminase domain sequence with hmmscan (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan). The resulting domain sequences were then used for structure classification. Because the core deaminase domain used for clustering may not show editing activity, we synthesized some of the candidate deaminases with elongated N-terminal and C-terminal sequences from each corresponding gene. This extension will help to enhance protein stability and ensure the deaminase activity can be fully played, and then we evaluated their cytidine deaminase activity with the reporter system or at endogenous sites.

(B) Protein sequence-based inference of the phylogeny of the members of the CDA superfamily. Except for the JAB superfamily as outgroup, different cytidine deaminase families are shown by different color modes. Nodes with Bootstrap ≥90 are identified by circles.

(C) Clustering results of structures and sequences based on different methods. The results of family classification based on the InterPro database were used as reference. The top and bottom was the different clustering results of structures and sequences, respectively. The default maximum cluster components are set to 21.

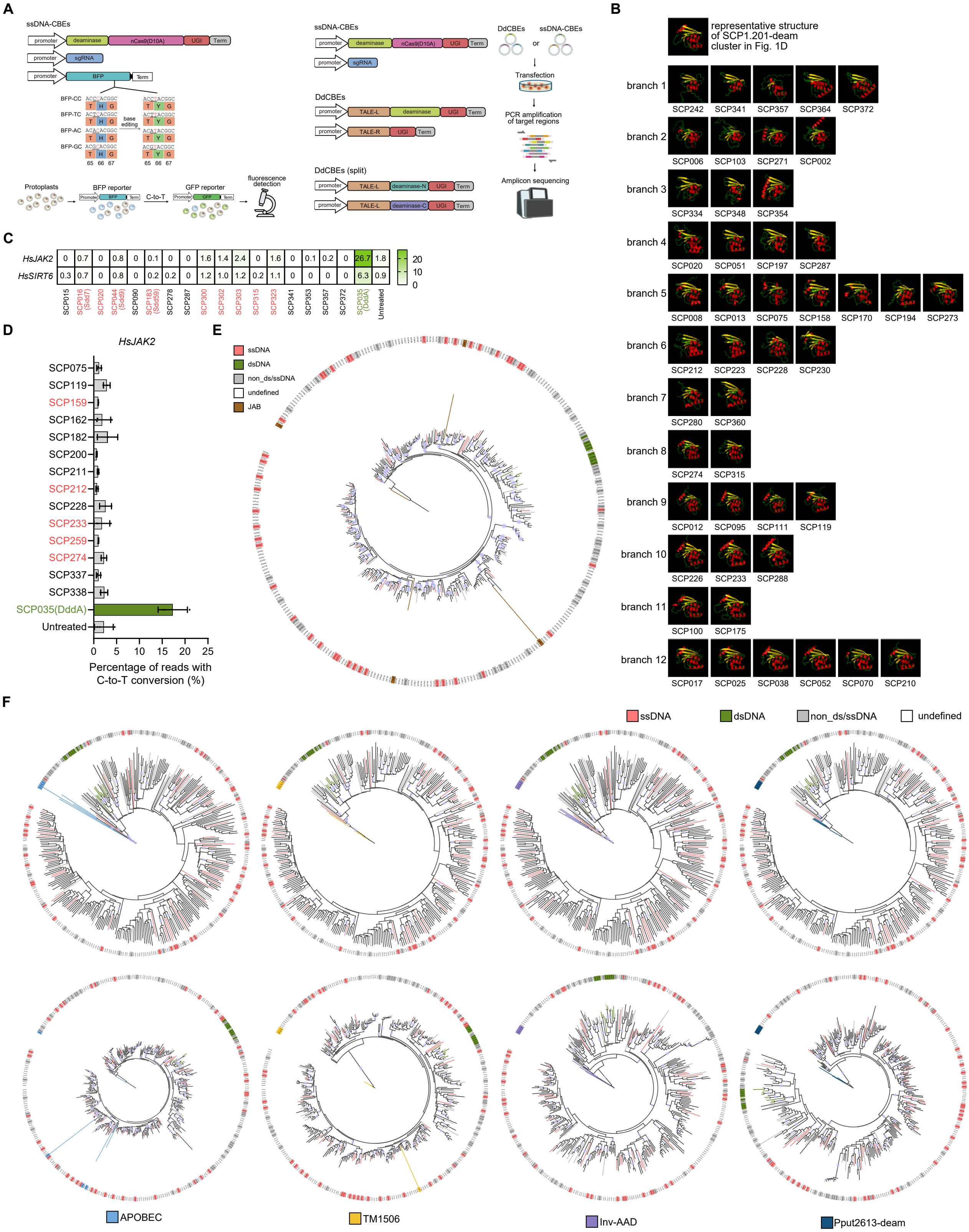
(D) Alignment of representative structures of the separate LmjF365940, APOBEC, dCMP, and MafB19 clades corresponding to Figure 1B. The two represented structures of each clade are also shown with pLDDT. Although the pairs of clades from each of the four families had partially similar structures, their overall structures displayed relatively large differences, leading them to be classified as different.

(E) Protein structural alignment of 15 SCP1.201 proteins.

(F) Protein structural alignment of 15 Toxin_deam proteins.

(G) Protein structural alignment between 15 SCP1.201 proteins and 1 Toxin_deam protein.

(H) Protein structural alignment between 15 Toxin_deam proteins and 1 SCP1.201 protein.



(legend on next page)

Figure S2. Validation of the diverse functions of SCP1.201 deaminases, related to Figure 2

(A) Left, reporter system for ssDNA cytidine deamination activity identification. Top of the left panel, schematic diagram of the ssDNA base editing vector for the BFP reporter system. Bottom of the left panel, the procedure used to detect Sdds catalyzing C-to-T changes using the BFP reporter system in rice protoplasts. Right, ssDNA and dsDNA cytidine deamination activity identification at endogenous sites. Left of the right panel, schematic diagram of the ssDNA base editing vector for the endogenous site editing and the DdCBEs vector and its split form. Right of the right panel, the procedure used to detect the activity of DdCBEs on dsDNA as well as ssDNA CBEs on ssDNA in HEK293T cells, respectively, followed by high-throughput sequencing.

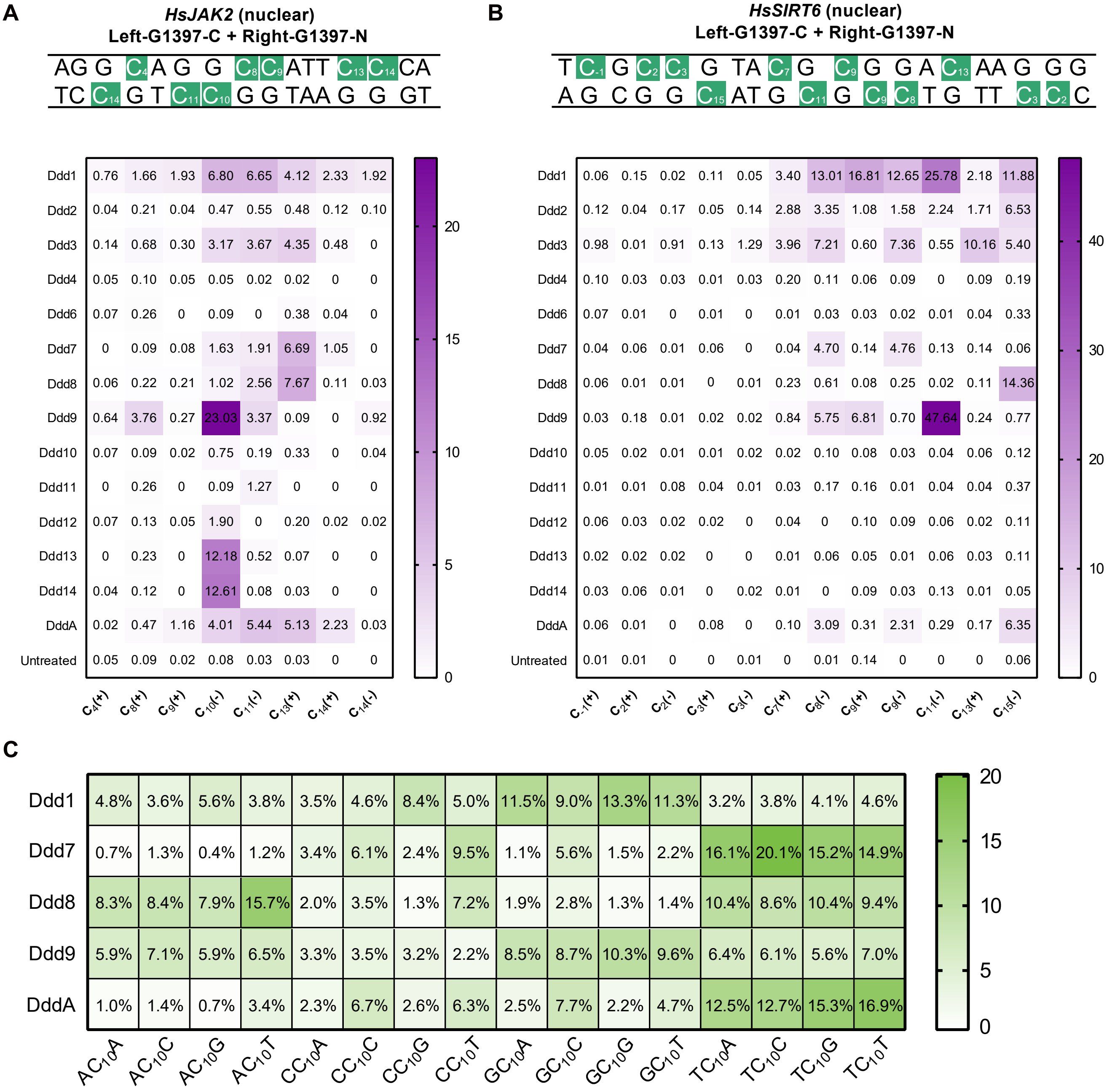
(B) Comparison between the representative structure of SCP1.201 in Figure 1C and the representative structure of each branch of the 12 branches in Figure 2A. The representative structures of SCP1.201 as shown in Figure 1C and all representative structures from 12 branches of the entire SCP1.201 structural tree according to the order from left to right on the SCP1.201 structural tree. Branch 2 is the Ddd cluster.

(C) Heatmap summarizing the editing efficiencies of dsDNA substrates of Sdds at *HsJAK2* and *HsSIRT6* sites. The gene name of active Sdd is colored in red. DddA with dsDNA deamination activity is colored in green. Data are representative of three independent experiments.

(D) Editing efficiencies of dsDNA substrates of some non-Ddd proteins at *HsJAK2* sites. The gene name of active Sdd is colored in red. DddA with dsDNA deamination activity is colored in green. Dots represent individual biological replicates, bars represent mean values, and error bars represent the SD of three independent biological replicates.

(E) Protein sequence-based inference of the phylogeny of the members of the SCP1.201 family. The JAB families are colored brown and regarded as an outgroup, and the tested deaminases are shown in red, green, and dark gray. Undefined deaminases in light gray await further functional analysis. Nodes with Bootstrap ≥90 are identified by circles.

(F) Protein sequence-based inference of the phylogeny of the members of the SCP1.201 family using four outgroups that more closely relate to the SCP1.201 family. The first row is based on structural classification, while the second row is based on sequence classification. Nodes with Bootstrap ≥90 are identified by circles.



	C_{4}(+)	C_{8}(+)	C_{9}(+)	C_{10}(-)	C_{11}(-)	C_{13}(+)	C_{14}(+)	C_{14}(-)
Ddd1	0.76	1.66	1.93	6.80	6.65	4.12	2.33	1.92
Ddd2	0.04	0.21	0.04	0.47	0.55	0.48	0.12	0.10
Ddd3	0.14	0.68	0.30	3.17	3.67	4.35	0.48	0
Ddd4	0.05	0.10	0.05	0.05	0.02	0.02	0	0
Ddd6	0.07	0.26	0	0.09	0	0.38	0.04	0
Ddd7	0	0.09	0.08	1.63	1.91	6.69	1.05	0
Ddd8	0.06	0.22	0.21	1.02	2.56	7.67	0.11	0.03
Ddd9	0.64	3.76	0.27	23.03	3.37	0.09	0	0.92
Ddd10	0.07	0.09	0.02	0.75	0.19	0.33	0	0.04
Ddd11	0	0.26	0	0.09	1.27	0	0	0
Ddd12	0.07	0.13	0.05	1.90	0	0.20	0.02	0.02
Ddd13	0	0.23	0	12.18	0.52	0.07	0	0
Ddd14	0.04	0.12	0	12.61	0.08	0.03	0	0
DddA	0.02	0.47	1.16	4.01	5.44	5.13	2.23	0.03
Untreated	0.05	0.09	0.02	0.08	0.03	0.03	0	0

	C_{-1}(+)	C_{2}(+)	C_{2}(-)	C_{3}(+)	C_{3}(-)	C_{7}(+)	C_{8}(-)	C_{9}(+)	C_{9}(-)	C_{11}(-)	C_{13}(+)	C_{15}(-)
Ddd1	0.06	0.15	0.02	0.11	0.05	3.40	13.01	16.81	12.65	25.78	2.18	11.88
Ddd2	0.12	0.04	0.17	0.05	0.14	2.88	3.35	1.08	1.58	2.24	1.71	6.53
Ddd3	0.98	0.01	0.91	0.13	1.29	3.96	7.21	0.60	7.36	0.55	10.16	5.40
Ddd4	0.10	0.03	0.03	0.01	0.03	0.20	0.11	0.06	0.09	0	0.09	0.19
Ddd6	0.07	0.01	0	0.01	0	0.01	0.03	0.03	0.02	0.01	0.04	0.33
Ddd7	0.04	0.06	0.01	0.06	0	0.04	4.70	0.14	4.76	0.13	0.14	0.06
Ddd8	0.06	0.01	0.01	0	0.01	0.23	0.61	0.08	0.25	0.02	0.11	14.36
Ddd9	0.03	0.18	0.01	0.02	0.02	0.84	5.75	6.81	0.70	47.64	0.24	0.77
Ddd10	0.05	0.02	0.01	0.01	0.02	0.02	0.10	0.08	0.03	0.04	0.06	0.12
Ddd11	0.01	0.01	0.08	0.04	0.01	0.03	0.17	0.16	0.01	0.04	0.04	0.37
Ddd12	0.06	0.03	0.02	0.02	0	0.04	0	0.10	0.09	0.06	0.02	0.11
Ddd13	0.02	0.02	0.02	0	0	0.01	0.06	0.05	0.01	0.06	0.03	0.11
Ddd14	0.03	0.06	0.01	0	0.02	0.01	0	0.09	0.03	0.13	0.01	0.05
DddA	0.06	0.01	0	0.08	0	0.10	3.09	0.31	2.31	0.29	0.17	6.35
Untreated	0.01	0.01	0	0	0.01	0	0.01	0.14	0	0	0	0.06

	$AC_{10}A$	$AC_{10}C$	$AC_{10}G$	$AC_{10}T$	$CC_{10}A$	$CC_{10}C$	$CC_{10}G$	$CC_{10}T$	$GC_{10}A$	$GC_{10}C$	$GC_{10}G$	$GC_{10}T$	$TC_{10}A$	$TC_{10}C$	$TC_{10}G$	$TC_{10}T$
Ddd1	4.8%	3.6%	5.6%	3.8%	3.5%	4.6%	8.4%	5.0%	11.5%	9.0%	13.3%	11.3%	3.2%	3.8%	4.1%	4.6%
Ddd7	0.7%	1.3%	0.4%	1.2%	3.4%	6.1%	2.4%	9.5%	1.1%	5.6%	1.5%	2.2%	16.1%	20.1%	15.2%	14.9%
Ddd8	8.3%	8.4%	7.9%	15.7%	2.0%	3.5%	1.3%	7.2%	1.9%	2.8%	1.3%	1.4%	10.4%	8.6%	10.4%	9.4%
Ddd9	5.9%	7.1%	5.9%	6.5%	3.3%	3.5%	3.2%	2.2%	8.5%	8.7%	10.3%	9.6%	6.4%	6.1%	5.6%	7.0%
DddA	1.0%	1.4%	0.7%	3.4%	2.3%	6.7%	2.6%	6.3%	2.5%	7.7%	2.2%	4.7%	12.5%	12.7%	15.3%	16.9%

Figure S3. Evaluating the activities and properties of newly discovered Ddd proteins for use as DdCBEs, related to Figure 3

(A) Heatmap of editing efficiencies and editing windows of SCP1.201 dsDNA deaminases at *HsJAK2* target sites in HEK293T cells.

(B) Heatmap of editing efficiencies and editing windows of SCP1.201 dsDNA deaminases at *HsSIRT6* target sites in HEK293T cells.

(C) The proportion of editing efficiencies of each context preference among 16 plasmid libraries of different Ddds. Data are represented by the average of three independent experiments (n = 3).

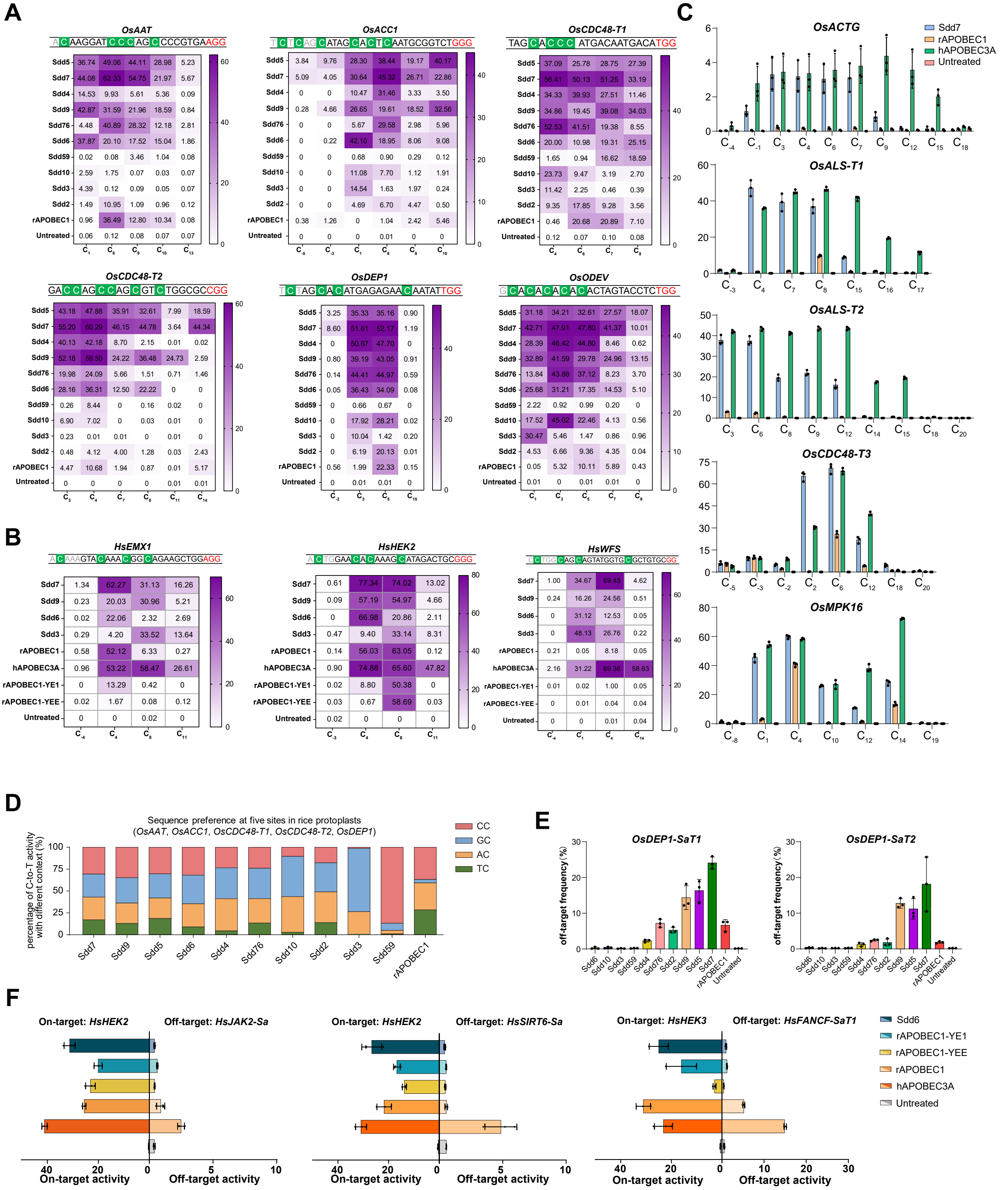


Figure S4. Evaluating the activities and properties of newly discovered Sdd proteins for use as base editors, related to Figure 4

(A) Editing behavior of Sdds and rAPOBEC1 at six endogenous target sites in rice protoplasts. The heatmap shows the editing efficiencies and editing windows of 10 Sdds and rAPOBEC1 at *OsAAT*, *OsACC1*, *OsCDC48-T1*, *OsCDC48-T2*, *OsDEP1*, and *OsODEV* sites in rice protoplasts. The values given in the heatmap cells

(legend continued on next page)

represent C-to-T editing efficiencies. Target sequences are listed above the heatmap, with green boxes marking the positions of C-to-T edits and protospacer adjacent motifs (PAMs) in red font. Data are represented by the average of three independent experiments.

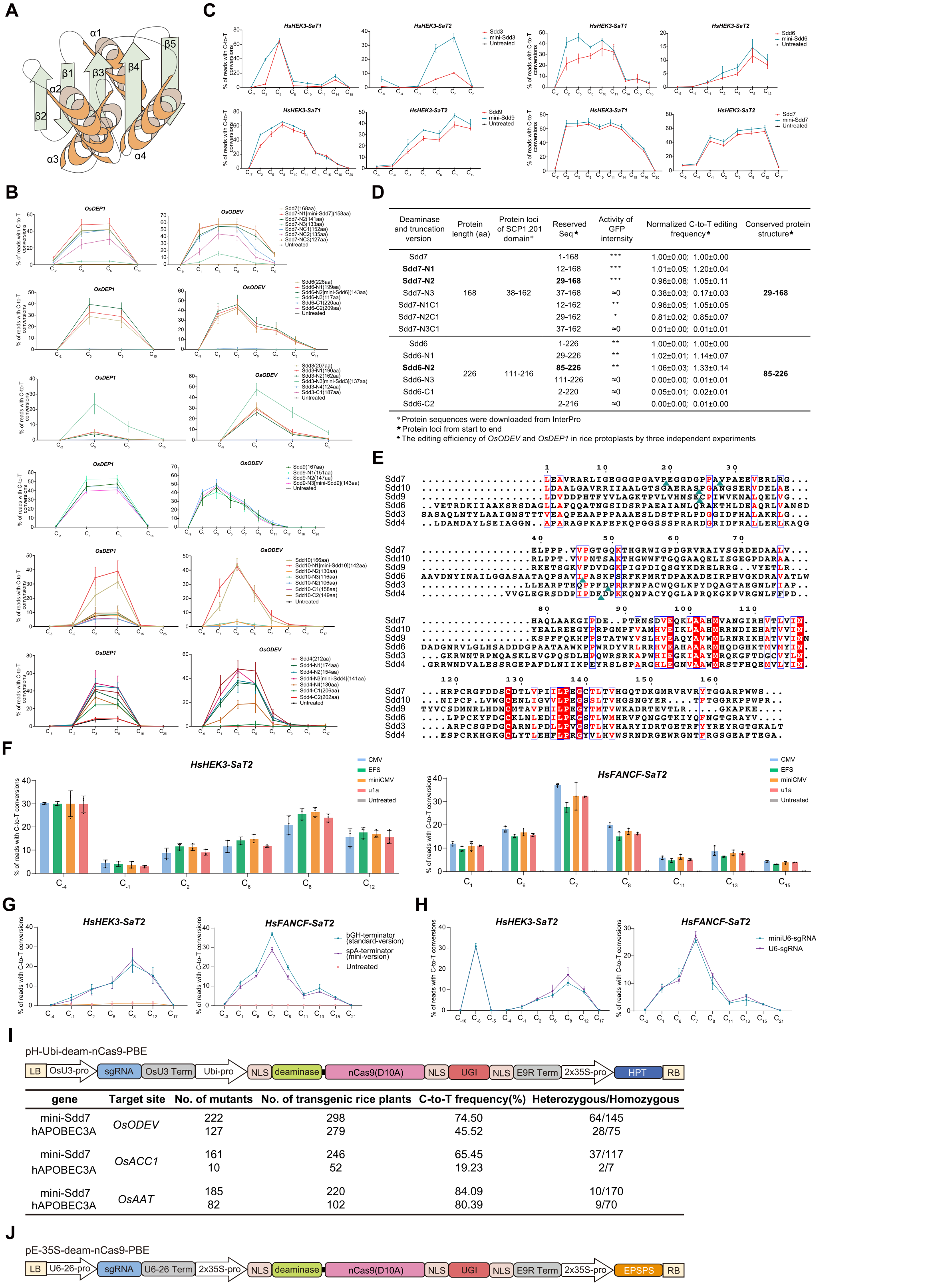
(B) Editing behavior of SCP1.201 ssDNA deaminases and APOBECs at three endogenous target sites in HEK293T cells. The heatmap gives the editing efficiencies and editing windows of four Sdds, rAPOBEC1, hA3A, rAPOBEC1-YE1, and rAPOBEC1-YEE at the *HsEMX1*, *HsHEK2*, *HsWFS1* sites in HEK293T cells. The values given in the heatmap cells represent C-to-T editing efficiencies. Target sequences are listed above the heatmap, with green boxes marking the positions of C-to-T edits and PAMs in red font. Data are represented by the average of three independent experiments.

(C) Comparison of the efficiencies of Sdd7, rAPOBEC1, and hAPOBEC3A at five sites in rice protoplasts. The efficiencies of Sdd7, rAPOBEC1, and hAPOBEC3A base editors compared across five endogenous targets, *OsACTG*, *OsALS-T1*, *OsALS-T2*, *OsCDC48-T3*, and *OsMPK16*. Dots represent individual biological replicates, bars represent mean values, and error bars represent the SD of three independent biological replicates.

(D) Sequence preference of Sdds and rAPOBEC1 at five endogenous target sites in rice protoplasts. The stacked graph shows the context preferences of 10 Sdds and rAPOBEC1 at five endogenous target sites, *OsAAT*, *OsACC1*, *OsCDC48-T1*, *OsCDC48-T2*, and *OsDEP1*. The green, yellow, blue, and red bars represent the proportions of C-to-T activity for TC, AC, GC, and CC, respectively. Data are representative of three independent experiments.

(E) Frequencies of off-target events for Sdds and rAPOBEC1 at two endogenous target sites in rice protoplasts. Off-target events were evaluated using the orthogonal R-loop assay. Frequencies of off-target events for Sdds and rAPOBEC1 at the *OsDEP1-SaT1* and *OsDEP1-SaT2* sites in rice protoplasts. Dots represent individual biological replicates, bars represent mean values, and error bars represent the SD of three independent biological replicates.

(F) On-target and off-target editing efficiency of Sdd6 and APOBEC base editors tested across two on-target and three off-target sites in HEK293T cells. Detailed display of on-target and off-target activity. On-target and off-target editing efficiencies of Sdd6, rAPOBEC1-YE1, rAPOBEC1-YEE, rAPOBEC1, and hAPOBEC3A across the *HsHEK2* and *HsHEK3* on-target sites and the *HsJAK2-Sa*, *HsSIRT6-Sa*, *HsRNF2-Sa*, and *HsFANCF-SaT1* off-target sites. Dots represent individual biological replicates, bars represent mean values, and error bars represent the SD of three independent biological replicates.



Deaminase and truncation version	Protein length (aa)	Protein loci of SCP1.201 domain*	Reserved Seq★	Activity of GFP internsity	Normalized C-to-T editing frequency♣	Conserved protein structure★
Sdd7	168	38-162	1-168	***	1.00±0.00; 1.00±0.00	**29-168**
Sdd7-N1			12-168	***	1.01±0.05; 1.20±0.04	
Sdd7-N2			**29-168**	***	0.96±0.08; 1.05±0.11	
Sdd7-N3			37-168	≈0	0.38±0.03; 0.17±0.03	
Sdd7-N1C1			12-162	**	0.96±0.05; 1.05±0.05	
Sdd7-N2C1			29-162	*	0.81±0.02; 0.85±0.07	
Sdd7-N3C1			37-162	≈0	0.01±0.00; 0.01±0.01	
Sdd6	226	111-216	1-226	**	1.00±0.00; 1.00±0.00	**85-226**
Sdd6-N1			29-226	**	1.02±0.01; 1.14±0.07	
Sdd6-N2			**85-226**	**	1.06±0.03; 1.33±0.14	
Sdd6-N3			111-226	≈0	0.00±0.00; 0.01±0.01	
Sdd6-C1			2-220	≈0	0.05±0.01; 0.02±0.01	
Sdd6-C2			2-216	≈0	0.00±0.00; 0.01±0.00	

* Protein sequences were downloaded from InterPro
★ Protein loci from start to end
♣ The editing efficiency of *OsODEV* and *OsDEP1* in rice protoplasts by three independent experiments

```
                              1         10         20        30
Sdd7   ....................LEAVRARLIGEGGGPGAVPEGGDGPPAVPAEEVERLRG...
Sdd10  ....................LDAALGAVRRIIAALGTSGAERASPGANGSERVDELAE...
Sdd9   ....................LDVDDPHTFYVLAGKTPVLVHNSECPIWVKNALQELVG...
Sdd6   ..VETRDKIIAAKSRSDAGLLAFQQATNGSIDSRPAEAIANLQRAKTHLDEAQRLVANSD
Sdd3   SASAQLNTYLAAIGNSTTTVEAQPEAAPPPAAAESLDSTPRLPDGGIDFHALAKRLG...
Sdd4   ..LDAMDAYLSEIAGGN..APARAGPKAPEPKQPGGSSSPRARDGRIDFRALLERLKAQG

                          40         50         60         70
Sdd7   ....................ELPPP.VVPGTGQKTHGRWIGPDGRVRAIVSGRDEDAALV...
Sdd10  ....................RLPPT.VVPNTSAKTHGWWFTGQGAAQELISGEGPDARAA...
Sdd9   ....................RKETSGKVFDVDGKPIGPDIQSGYKDRELRRG..VYETLR...
Sdd6   AAVDNYINAILGGASAATAQPSAVIPASKPSRFKPMRTDPAKADEIRPHVGKDRAVATLW
Sdd3   ...................LLEARPTEQPPFDPRRFNPACWQGLKPYDQAGTAEGNLFIAP..
Sdd4   ..............VVGLEGRSDDPIPDFDPKKQNPACYQGLAPRQKGKPVRGNLFFPD..

                          80         90        100        110
Sdd7   ....................HAQLAAKGIPDE..PTRNSDVEQKLAAHMVANGIRHVTLVIN.
Sdd10  ....................YEALREEGYPRPGMPFVAMHVEIKLAAHMRRNDIEHATVVIN.
Sdd9   ....................KSPFFQKHFPSTATWYVSLHVEAQYAVWMLRNRIKHATVVINN
Sdd6   DADGNRVLGLHSADDDGPAATAAWKPPWRDYVRLRRHVEAHAAARMHQDGHKTMVMYIN.
Sdd3   ...GKRWNTRPMQASKLEVGPQSDLHPQWRSRKAPWHIEGKIAAYMRQKGFTDGCVYLN.
Sdd4   .GRRWNDVALESSRGEPAFDLNIIKPEYRSLSPARGHLEGNVAAWMRSTFHQEMVLYIN.

          120        130        140        150        160
Sdd7   ...HRPCRGFDDSCDTLVPIILPEGCTLTVHGQTDKGMRVRVRYTGGARPWWS...
Sdd10  ...NIPCP.LVWGCENLIGVVLPEGSSLTVHGSNGYER....TFTGGRKPPWPR..
Sdd9   TYVCSDMNRLHDNCMTAVPHILPEGYTMTVWKADRTEVTLR......GKAPKE....
Sdd6   ....LPPCKYFDGCKLNLEDILPKGSTLWMHRVFQNGGTKIYQFNGTGRAYV....
Sdd3   ....ARPCSGPDGCARNLPDLLPVGSTLHVHARYIDRTGETRFYYREYRGTGKALT
Sdd4   ...ESPCRKHGKGCLYTLEHFLPRGYVLHVWSRNDRGEWRGNTFRGSGEAFTEGA.
```

gene	Target site	No. of mutants	No. of transgenic rice plants	C-to-T frequency(%)	Heterozygous/Homozygous
mini-Sdd7 hAPOBEC3A	*OsODEV*	222 127	298 279	74.50 45.52	64/145 28/75
mini-Sdd7 hAPOBEC3A	*OsACC1*	161 10	246 52	65.45 19.23	37/117 2/7
mini-Sdd7 hAPOBEC3A	*OsAAT*	185 82	220 102	84.09 80.39	10/170 9/70

(legend on next page)

Figure S5. Optimization and development of Sdd-CBEs for therapeutic and agricultural applications, related to Figure 5

(A) Conserved protein structure of Sdds with high activity predicted by AlphaFold2. The core structure of Sdds with high deamination activity is shown. For some deaminases, $\alpha4$ is not essential.

(B) Testing of the efficiencies of different truncated versions of synthesized deaminase genes at two endogenous target sites in rice protoplasts. Removal of redundant sequence of synthesized deaminase genes assisted by AlphaFold2. Editing efficiencies of multiple redundant sequence removal versions of Sdd7, Sdd6, Sdd3, Sdd9, Sdd10, and Sdd4 deaminases at the *OsDEP1* and *OsODEV* sites in rice protoplasts are shown. Truncations included various forms of C-terminal and N-terminal deletions. Data are representative of three independent experiments. Dots represent mean values, and error bars represent the SD of three independent biological replicates.

(C) Testing of the efficiencies of different truncated versions of synthesized deaminase genes at two endogenous target sites in human cells. Removal of redundant sequence of synthesized deaminase genes assisted by AlphaFold2. Comparison of the Sdd3, Sdd9, Sdd6, and Sdd7 deaminases and their redundant sequence removal versions at the *HsHEK3-SaT1* and *HsHEK3-SaT2* sites in HEK293T cells. Data are representative of three independent experiments. Dots represent mean values, and error bars represent the SD of three independent biological replicates.

(D) The intensity of GFP reporter fluorescence and endogenous editing efficiency of wild type and different truncated variants of Sdd7 and Sdd6 based on the predicted protein structures.

(E) The multiple protein sequence results of six high-activity Sdd proteins. The cyan arrows represent the highest activated truncated N-terminal sites of deaminases variants, including mini-Sdd7, Sdd7-N2, mini-Sdd10, mini-Sdd9, mini-Sdd6, mini-Sdd3, and mini-Sdd4.

(F) Comparison the editing efficiencies of mini-Sdd6 with four promoters, CMV, EFS, mini-CMV, and u1a at the *HsHEK3-SaT2* and *HsFANCF-SaT2* sites in HEK293T cells. Dots represent individual biological replicates, bars represent mean values, and error bars represent the SD of three independent biological replicates.

(G) Comparison of mini-Sdd6 with two terminators, bGH and SpA, at the *HsHEK3-SaT2* and *HsFANCF-SaT2* sites in HEK293T cells to observe the effects of terminators on AAV vectors. Dots represent mean values, and error bars represent the SD of three independent biological replicates.

(H) Comparison of minU6 and U6 promoters at the *HsHEK3-SaT2* and *HsFANCF-SaT2* sites in HEK293T cells. Data are representative of three independent experiments.

(I) Frequencies of base-edited regenerated rice plants. Top, schematic diagram of the base editing binary vector for *Agrobacterium*-mediated transformation in rice. Bottom, frequencies of mutations induced by mini-Sdd7 and hAPOBEC3A base editors in T_0 rice plants.

(J) Schematic diagram of the base editing binary vector for *Agrobacterium*-mediated transformation in soybean.