

1 **Discovery of new deaminase functions by structure-based protein
2 clustering**

3 Jiaying Huang^{1,8}, Qiupeng Lin^{1,8}, Hongyuan Fei^{1,2,8}, Zixin He^{1,2,8}, Hu Xu³, Yunjia Li^{1,2},
4 Kunli Qu⁴, Peng Han⁴, Qiang Gao³, Boshu Li^{1,2}, Guanwen Liu¹, Lixiao Zhang³,
5 Jiacheng Hu¹, Rui Zhang¹, Erwei Zuo⁵, Yonglun Luo⁴, Yidong Ran³, Jin-Long Qiu^{6,7},
6 Kevin Tianmeng Zhao^{3*}, Caixia Gao^{1,2,9*}

7

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

11 ² College of Advanced Agricultural Sciences, University of Chinese Academy of
12 Sciences, Beijing, China.

13 ³ Qi Biodesign, Beijing, China.

14 ⁴ Lars Bolund Institute of Regenerative Medicine, Qingdao-Europe Advanced Institute
15 for Life Sciences, BGI-Qingdao, Qingdao, China.

16 ⁵ Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome
17 Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at
18 Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China.

19 ⁶ State Key Laboratory of Plant Genomics, Institute of Microbiology, Chinese
20 Academy of Sciences, Beijing, China.

21 ⁷ CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of
22 Sciences, Beijing, China

23 ⁸ These authors contributed equally to this work.

24 ⁹ Lead contact

25 *Correspondence: kzhao@qi-biodesign.com, cxgao@genetics.ac.cn

26 **Summary**

27 The elucidation of protein function and its exploitation in bioengineering have greatly
28 contributed to the development of the life sciences. Existing protein mining efforts
29 generally rely on amino acid sequences rather than protein structures due to technical
30 difficulties in structural elucidation. We describe here for the use of AlphaFold2 to
31 predict and subsequently cluster an entire protein family based on predicted structure
32 similarities. We selected the deaminase family of proteins to analyze and through this
33 approach identified many previously unknown deaminase properties. We applied
34 these new deaminases to the development of new cytosine base editors with distinct
35 features. Although we found many new double-stranded DNA deaminases from the
36 DddA-like protein clade, we were surprised to find that most of the proteins in this
37 family were not actually double-stranded DNA cytidine deaminases. From this protein
38 clade, we engineered the smallest single-strand specific cytidine deaminase, which
39 facilitates the first efficient cytosine base editor to be packaged into a single AAV
40 vector. Importantly, we also profiled a deaminase from this clade that edits robustly in
41 soybean plants, which previously suffered from poor editing by cytosine base editors.
42 These newly discovered deaminases based on AI-assisted structural predictions
43 greatly expand the utility of base editors for therapeutic and agricultural applications.

44

45 Keywords: structural prediction, protein classification, deaminase, Ddd, Sdd,
46 specificity, context preference, base editing

47 **Introduction**

48 The discovery and engineering of new proteins has greatly transformed the life
49 sciences. Traditional enzyme mining based solely on sequence information has been
50 effective at classifying and predicting protein functions and evolutionary trajectory^{1,2}.
51 However, one-dimensional (1D) information, whether in the form of core amino acids,
52 specific motifs, overall amino acid sequence identity, or Hidden Markov Models
53 (HMM), cannot completely illuminate the functional characteristics of proteins.

54 In contrast, since protein function is ultimately determined by three dimensional
55 (3D) protein folds, understanding protein structures would provide reliable and
56 rational insights into protein function during the process of protein mining and
57 clustering classifications^{3,4}. Although the number of publicly reported protein
58 structures is increasing, it is minuscule compared to the number of new proteins
59 discovered based on amino acid sequences^{5,6}. Recently, many artificial intelligence
60 (AI) methods have been developed that use 1D amino acid sequences to accurately
61 predict high resolution 3D protein structures⁷⁻⁹. These protein structure prediction
62 methods should thus enable large-scale mining and classifications of proteins with
63 specific functions.

64 Deaminase-like proteins catalyze the deamination of nucleotides and bases in
65 nucleic acids. They play important roles in defense, mutation and nucleic acid
66 metabolism and other biological processes¹⁰⁻¹³ and have been recently exploited for
67 use in programmable DNA and RNA base editors¹⁴⁻¹⁶, a class of precise genome
68 editing technologies. Members of this family act as nucleotide deaminases and nucleic
69 acid deaminases, including adenosine, cytidine, cytosine and guanine deaminases, and
70 have the ability to act on single-stranded DNA (ssDNA)¹⁷, double-stranded DNA
71 (dsDNA)¹⁰, double-stranded RNA (dsRNA)¹⁸, transfer RNA (tRNA)¹⁹, free
72 nucleosides¹², and other deaminated nucleotide derivatives²⁰. The sporadic
73 distribution of deaminases and their rapid evolution due to positive selection often
74 confounds the relationships between the various protein families in phylogenetic
75 analyses based on sequence^{20,21}. Here, we performed new protein clustering

76 classifications on the greater deaminase family of proteins based on
77 AlphaFold2-predicted 3D structures.

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

84 Cytosine base editors (CBEs) use cytidine deaminases to catalyze C-to-U base
85 conversions, resulting in permanent C•G-to-T•A base edits in DNA^{14,15,22,23}. Base
86 editors have great potential in therapeutic genome editing, fundamental life sciences
87 research, and for breeding new elite traits into plants²⁴⁻²⁶. Previous DNA base editors
88 exploited the use of two types of cytidine deaminases acting on either ssDNA or
89 dsDNA^{10,14}. To date, only a few ssDNA-targeting APOBEC/AID-like deaminases and
90 one dsDNA-targeting deaminase (DddA) have been used to generate CBEs^{10,14,15,27-30}.
91 These deaminases remain limited to sequence context restrictions, low on-off target
92 editing ratios and large protein sizes, which makes their delivery by adeno-associated
93 virus (AAV) viral vectors difficult³¹. For unknown reasons, some species like soybean
94 plants, a staple agricultural crop grown all over the world, have suffered from poor
95 cytosine base editing since the technology was first introduced in 2016³². Thus, robust
96 and more efficient CBEs are still needed to further expand their utility. By generating
97 new protein classifications based on their predicted structures, we have developed a
98 suite of new ssDNA and dsDNA deaminases used for precision genome editing. We
99 highlight that enzyme mining based on structures predicted by AlphaFold2 is a simple,
100 flexible, and high-throughput method to classify and engineer proteins with unknown
101 functions.

102 **Results**

103 **Clustering and discovery of new cytidine deaminases via protein structures**

104 We hypothesized that the comparison and clustering of known or predicted protein
105 structures, given that the 3D structure of a protein ultimately determines its function,
106 could be an effective method for classifying deaminases into functional clades. Thus,
107 we employed a combination of AI-assisted protein structure prediction, structural
108 alignments, and clustering to generate new protein classification relationships among
109 deaminases (Figure 1A). We selected 238 protein sequences annotated as having a
110 deaminase domain from the InterPro database and 4 distant outgroup candidate
111 protein sequences from the JAB-domain family (Figure S1A). Specifically, we
112 randomly selected 15 candidates of at least 100 amino acids in length from each of the
113 16 deaminase families and used AlphaFold2 to predict their protein structures. We
114 conducted multiple structural alignments (MSTA) of all candidates using normalized
115 TM-scores as a guide³³. Based on the MSTA results, we generated candidate
116 similarity matrices reflecting the overall structural correlation between the proteins.
117 We then organized these similarity matrices into a structural dendrogram using the
118 average-linkage clustering algorithm (Figure 1B). The dendrogram clustered the 238
119 proteins into 20 unique structural clades and the deaminases within each clade had
120 distinct conserved protein structural domains (Figure 1C and 1D).

121 We found that accurate protein clustering classifications could be generated based
122 on protein structural alignments, even without the use of contextual information such
123 as conserved gene neighborhoods and domain architectures. When using
124 structure-based hierarchical clustering, different clades reflected unique structures,
125 implying distinct catalytic functions and properties (Figure 1D). Interestingly, we also
126 found that this structure-based clustering method was much more effective at sorting
127 for functional similarities than traditional 1D amino acid sequence-based clustering
128 approaches. For example, adenosine deaminases (A_deamin, PF02137 in InterPro
129 database), enzymes involved in purine metabolism, were split into different clades

130 when using amino acid sequence-based clustering methods but were all grouped
131 together into a single A_deamin-clade using our structure-based clustering approach
132 (Figure 1B, 1C and S1B). Additionally, four deaminase families (dCMP, MafB19,
133 LmjF365940 and APOBEC as annotated by InterPro) were each divided into two
134 separate clades when using structure-based clustering (Figure 1C and 1D).
135 Comparison of protein structures showed that the two clades for each of these four
136 deaminase families had quite different structures, contrary to what their InterPro
137 naming and sequenced-based classification might suggest (Figure 1D and S1C). In
138 summary, AI-assisted 3D protein structures provide reliable clustering results and
139 only require an amino acid sequence from the user, making it a convenient and
140 effective strategy for generating protein relationships.

141 **Evaluating diverse deaminase clades by fluorescence imaging**

142 CRISPR-based CBEs are precise genome editing technologies capable of generating
143 C•G-to-T•A substitutions in the genome of living cells. Because single-strand DNA
144 specific cytidine deaminases are an essential component of CBEs, we sought to
145 explore the deamination activity of each structure-based classified deaminase clade in
146 the context of DNA base editing. We evaluated a total of 190 deaminase domains by
147 selecting at least five proteins from each clade. Importantly, because the core
148 deaminase domain used for clustering may not show editing activity, we extended
149 each deaminase sequence to include additional secondary structures from each
150 corresponding gene around the deaminase domain (Figure S1A). For each of 190
151 newly annotated deaminases, we generated plant CBEs by fusing each candidate
152 domain-related sequence to the N-terminus of a Cas9 nickase (nCas9, D10A)
153 followed by an uracil-DNA glycosylase inhibitor (UGI)^{14,34}. We developed four
154 BFP-to-GFP reporter systems to reflect TC, CC, GC, and AC 5'-base deamination
155 preferences (Figure S2A). Each CBE was co-transformed with all four BFP-to-GFP
156 reporter plasmids into rice protoplasts and analyzed by fluorescent microscopy after
157 three days³⁴. We found that deaminases belonging to the SCP1.201 (PF14428),

158 XOO_2897 (PF14440), MafB19 (PF14437), Toxin-deaminase (PF14424), and
159 TM1506 (PF08973) clades possessed ssDNA cytidine deamination activity.
160 Interestingly, we noticed that some deaminase candidates displayed different sequence
161 preferences compared to the APOBEC/AID-like deaminases as evaluated using the
162 fluorescence reporter system. Therefore, we demonstrated that the use of 3D
163 structures for protein classification enabled the discovery of new functional
164 deaminase clusters for use in base editors, offering new opportunities for developing
165 enhanced and bespoke precise base editing tools.

166 **Validation of the diverse functions of SCP1.201 deaminases**

167 While evaluating deaminases from each clade, we were surprised to find that some
168 deaminases annotated from the SCP1.201 clade were capable of deaminating
169 single-stranded DNA substrates. These deaminases were previously named
170 Double-stranded DNA deaminase toxin A-like (DddA-like) deaminases in the
171 InterPro database (PF14428). The DddA deaminase was recently developed into a
172 CRISPR-free double-stranded DNA cytosine base editor (DdCBE) capable of
173 deaminating cytosine bases on double-stranded DNA¹⁰. Because of DddA, all proteins
174 in the SCP1.201 clade were also annotated as double-stranded DNA deaminases. To
175 re-analyze this SCP1.201 clade, we selected all 489 SCP1.201 deaminases from the
176 InterPro database. We also included seven additional proteins that were 35% to 50%
177 identical by Basic Local Alignment Search Tool (BLAST) with DddA but were
178 characterized separately in InterPro. After identity and coverage filtering, we
179 performed a new AI-assisted protein structure-based classification of 332 SCP1.201
180 deaminases. Structure clustering showed that the SCP1.201 deaminases clustered into
181 different clades with unique core structural motifs (Figure 2A-2E).

182 We found that DddA and ten other proteins clustered into one subclade of
183 SCP1.201. Upon analyzing the 3D predicted structures of all 11 proteins within this
184 subclade, we found that they shared a similar core structure to DddA. Given their
185 structural similarities to DddA, we hypothesized that the other proteins in this

186 subclade can also perform double-strand DNA cytidine deamination. To evaluate
187 dsDNA deamination, we generated DdCBEs comprised of each deaminase alone or
188 split in half at a residue similar to the site where DddA was split by protein structure
189 alignment and joined together using a dual TALE system¹⁰ (Figure S2B). We
190 evaluated 10 proteins from this Ddd subclade in HEK293T cells at the *JAK2* and
191 *SIRT6* sites and observed that 8 proteins could perform dsDNA base editing (Figure
192 2A and 2F). We hereafter named these deaminases as double-strand DNA deaminases
193 (Ddd) and assigned them to this newly identified Ddd sub-clade.

194 To evaluate other SCP1.201 candidate proteins, we selected 24 proteins at
195 random and subjected these to our CBE fluorescent reporter system. We found that 22
196 showed detectable fluorescence and selected 13 to evaluate endogenous base editing
197 in the context of CBE in mammalian cells (Figure 2). Although these were previously
198 characterized as DddA-like, many showed cytosine base editing activity on ssDNA
199 (Figure 2A, 2G) but not dsDNA (Figure S2C). Therefore, we hereafter named these
200 ssDNA-targeting protein domains from the SCP1.201 clade as single-stranded DNA
201 deaminases (Sdd). We were surprised to find that a majority of protein members from
202 the SCP1.201 clade were found to be Sdd proteins since these were all previously
203 annotated as DddA-like. We also observed that these Sdd proteins shared a similar
204 protein structure as Sdd7, one of the highest editing ssDNA CBEs, which is distinct
205 from the Ddd proteins (Figure 2D and 2E). Thus, the annotated DddA-like
206 deaminases in the InterPro database (PF14428) should be further subdivided and
207 re-annotated accordingly.

208 In comparison, we also performed a clustering of the proteins from the SCP1.201
209 clade based on 1D amino acid sequences and found that some outgroup members
210 were dispersed throughout the tree, even though we chose four more closely related
211 families as outgroups (Figure S2D and S2E). These results highlight the usefulness
212 and importance of using protein structure-based classifications for comparing and
213 evaluating protein functional relationships.

214 **New Ddd proteins have distinct editing preferences to DddA**

215 Due to the strict 5'-TC sequence motif preference of DddA, the use of DddA-based
216 dsDNA base editors is limited predominantly to TC targets¹⁰. Although the recently
217 evolved DddA11 displayed a broadened ability to deaminate and edit cytosine bases
218 with a 5'-HC (H = A, C or T) motif, the editing efficiency for AC, CC, and GC targets
219 still need to be improved³⁵. We evaluated the newly discovered Ddd proteins to
220 determine if they could expand the utility and targeting scope of DdCBEs. 13
221 deaminases belonging to the Ddd sub-clade were cloned into DdCBEs and evaluated
222 for dsDNA base editing at the endogenous *JAK2* and *SIRT6* sites in HEK293T cells
223 (Figure 2F, S3A, S3B). Interestingly, we found that Ddd1, Ddd7, Ddd8, and Ddd9 had
224 comparable or higher editing efficiencies to DddA (Figure 3A, S3A and S3B).
225 Importantly, we identified that Ddd1 and Ddd9 had a much higher editing activity
226 compared to DddA at 5'-GC motifs (Figure 3A, S3A and S3B). Strikingly, at the C₁₀
227 (5'-GC) residue in *JAK2* and the C₁₁ (5'-GC) residue in *SIRT6*, we found that while
228 DddA resulted in 21.1% and 0.6% editing, Ddd9 was capable of editing 65.7% and
229 45.7%, respectively (Figure 3A).

230 Because certain Ddd proteins seemed to exhibit distinct editing patterns
231 compared to DddA, we sought to evaluate any sequence motif preference for these
232 Ddd proteins. We first constructed 16 plasmids³⁵ encoding the *JAK2* target sequence
233 and modified positions 9-11 from GCC to NCN (N = A, T, C and G), yielding 16
234 different plasmids, and independently co-transfected each plasmid along with a
235 DdCBE variant (Figure 3B). Following comparative analyses of C•G-to-T•A base
236 conversion frequencies for each NCN, we generated corresponding sequence motif
237 logos to reflect sequence context preferences of each dsDNA deaminase (Figure 3B).
238 We found that as previously discussed, DddA and its structural homolog, Ddd7,
239 strongly preferred a 5'-TC sequence motif (Figure 3C and S3C). In contrast, we found
240 that Ddd1 and Ddd9 showed preferences towards editing 5'-GC substrates, while
241 Ddd8 showed preferences towards editing 5'-WC (W=A or T) substrates. Therefore,
242 the newly discovered dsDNA-targeting deaminases can edit cytosine bases at motifs

243 previous inaccessible to DddA, which is also essential for future engineering efforts.

244 **Sdd deaminases enable base editing in human cells and plants**

245 We next wondered whether the newly characterized Sdd proteins could be used for
246 more precise or efficient base editing. We chose to evaluate the six most active Sdds
247 as well as four weaker Sdds and compared their activities using a fluorescent reporter
248 system. We generated plant CBEs for each of the ten Sdds and evaluated their
249 endogenous base editing across six sites in rice protoplasts (Figure 4A and S4A). We
250 found that seven of the deaminases (Sdd7, Sdd9, Sdd5, Sdd6, Sdd4, Sdd76 and Sdd10)
251 had higher activity compared to the rat APOBEC1 (rAPOBEC1)-based CBE. The
252 most active Sdd7 base editor reached as high as 55.6% cytosine base editing, which
253 was more than 3.5-fold that of rAPOBEC1. To examine the versatility of these
254 deaminases, we also constructed the corresponding human-cell targeting BE4max
255 vectors³⁶ and evaluated their editing efficiencies across three endogenous target sites
256 in HEK293T cells. In agreement with the results in rice, we found that Sdd7 had the
257 highest editing activity (Figure S4B).

258 We previously showed that human APOBEC3A (hA3A) performed robust base
259 editing with a large editing window in plants^{37,38}. We therefore compared the editing
260 activities of hA3A and Sdd7 in human cells (Figure S4B) and plants (Figure S4C).
261 Interestingly, Sdd7 had comparable editing activities as hA3A across all three target
262 sites in HEK293T cells (Figure S4B) and five endogenous sites in rice protoplasts
263 (Figure S4C). Because editing efficiency is of primary significance for genome
264 editing in plant breeding, these results confirmed that Sdd7 is a robust cytosine base
265 editor for use in both plants and human cells.

266 **Sdd proteins have unique base editing characteristics**

267 When evaluating endogenous base editing, we observed different editing patterns by
268 the different Sdd-CBEs across all tested genomic target sites in both human and rice
269 cells. For instance, while Sdd7, Sdd9, and Sdd6 showed no particular motif editing

270 preference, Sdd3 seemed to prefer editing 5'-GC and 5'-AC motifs and strongly
271 disfavor editing 5'-TC and 5'-CC motifs (Figure S4D). To better profile the editing
272 patterns of each deaminase, we used Targeted Reporter Anchored Positional
273 Sequencing (TRAP-seq), a high-throughput approach for parallel quantification of
274 base editing outcomes³⁹. A 12K TRAP-seq library comprised of 12,000 TRAP
275 constructs, each containing a unique gRNA expression cassette and the corresponding
276 surrogate target site, was stably integrated into HEK293T cells by lentiviral
277 transduction. Following cell culture and antibody selection, base editors were stably
278 transfected into this 12K-TRAP cell line followed by ten days of blasticidin selection
279 (Figure 4B). On the eleventh day post transfection, we extracted the genomic DNA
280 and performed deep amplicon sequencing to evaluate the editing products of each
281 deaminase (Figure 4B). We found that while Sdd7 and Sdd6 showed no strong
282 sequence context preference, rAPOBEC1 had a strong preference for 5'-TC and
283 5'-CC bases while disfavoring 5'-GC and 5'-AC bases (Figure 4C). On the contrary,
284 Sdd3 showed an entirely complementary pattern preferring to edit 5'-GC and 5'-AC
285 bases while showing nearly no activity towards 5'-TC and 5'-CC bases (Figure 4C).
286 Interestingly, we found that Sdd6 and Sdd3 had different editing windows and
287 preferred to edit positions +1 to +3 in the protospacer as compared to rAPOBEC1 and
288 Sdd7 (Figure 4C). In conclusion, the newly identified Sdd base editors show unique
289 base editing properties such as increased editing efficiencies, disparate deamination
290 motif preferences, and altered editing windows from conventional cytosine base
291 editors.

292 It was previously described that CBEs could cause genome-wide
293 Cas9-independent off-target editing outcomes, which raises concerns about the safety
294 of these precise genome editing technologies for clinical applications^{40,41}. It is thought
295 that these off-target mutations may be a result of overexpression of the cytidine
296 deaminase. We wondered whether the newly-discovered Sdd proteins could offer a
297 more favorable balance between off-target and on-target editing. We therefore
298 evaluated the Cas9-independent off-target effects of the ten Sdds using an established
299 orthogonal R-loop assay in rice protoplasts⁴². We found that six (Sdd2, Sdd3, Sdd4,

300 Sdd6, Sdd10, and Sdd59) of the ten deaminases had lower off-target activities than
301 rAPOBEC1. Interestingly, while Sdd6 showed nearly no off-target editing activity, it
302 was still robust at on-target base editing when tested across six endogenous sites in
303 rice protoplasts (Figure 4D and S4E). When we analyzed the on-target:off-target
304 ratios of these ten deaminases, Sdd6 exhibited the highest on-target:off-target editing
305 ratios, which was 37.6-fold that of rAPOBEC1 (Figure 4E). We further compared the
306 on-target and off-target editing of Sdd6 to that of rAPOBEC1 and its two high-fidelity
307 deaminase variants, YE1 and YEE, in HEK293T cells⁴³. Importantly, we found that
308 Sdd6 had the highest on-target:off-target editing ratios and was calculated to be
309 2.8-fold, 2.1-fold and 2.5-fold higher than that of rAPOBEC1, YE1 and YEE,
310 respectively, and 10.4-fold higher than that of hA3A (Figure 4F and S4F). Notably,
311 the on-target activity of Sdd6 was comparable to that of rAPOBEC1 and much higher
312 than that of YE1 and YEE (Figure S4F). Thus, we identified that the SCP1.201 clade
313 contains unique and more precise Sdd proteins to be used as high-fidelity base editors.

314 **Rational design of Sdd proteins assisted by AlphaFold2 structure prediction**

315 Although viral delivery of CBEs has great potential for disease treatment, the large
316 size of APOBEC/AID-like deaminases restricts their ability to be packaged into single
317 AAV particles for *in vivo* editing applications³¹. Others have developed dual-AAV
318 strategies delivery approaches by splitting CBEs into an amino-terminal and
319 carboxy-terminal fragment and packaging them into separate AAV particles³¹.
320 However, these delivery efforts would challenge large-scale manufacturing, require
321 higher viral dosages, and pose potential safety challenges for human use⁴⁴. Recently, a
322 truncated sea lamprey cytidine deaminase-like 1 (PmCDA1)-based CBE was
323 developed that could theoretically be packaged into a single-AAV, but the editing
324 efficiency was extremely low when using the packaged AAVs during HEK293T cell
325 transduction⁴⁵. As SCP1.201 deaminases are canonically compact and conserved
326 (Figure S5A), we thought that they might be the ideal protein for single-AAV CBEs.

327 We wondered whether we could use AI-assisted protein modeling to further

328 engineer and shorten the size of the newly discovered Sdd proteins. We then
329 generated multiple truncated variants of each of Sdd7, Sdd6, Sdd3, Sdd9, Sdd10, and
330 Sdd4 and tested these variants for endogenous base editing in rice protoplasts across
331 two sites each.

332 We identified mini-Sdd7, mini-Sdd6, mini-Sdd3, mini-Sdd9, mini-Sdd10, and
333 mini-Sdd4 as newly minimized deaminases that are both small (~130-160 aa) and
334 have comparable or higher editing efficiencies compared to their full-length proteins
335 both in rice protoplasts and human cells (Figure 5A, S5B and S5C). Strikingly, all six
336 miniaturized deaminases would permit the construction of single-AAV-encapsulated
337 SaCas9-based CBEs (< 4.7 kb between ITRs) (Figure 5B, S5D, S5E and S5F). We
338 used mini-Sdd6 to construct a single-AAV SaCas9 vector and found that it had editing
339 efficiencies of around 60% in mouse neuroblastoma N2a cells at two sites in the *HPD*
340 gene (*Mus musculus* 4-hydroxyphenylpyruvate dioxygenase)⁴⁶ by transient
341 transfection (Figure 5C). These results highlight that the Sdd proteins offer great
342 advantages over APOBEC/AID deaminases in terms of AAV-based CRISPR base
343 editing delivery. The success in further shortening Sdd proteins for AAV packaging
344 highlights the great potential of AI-assisted protein engineering.

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

346 We next explored the use and application of newly engineered Sdd proteins for base
347 editing in plants. We first evaluated the ability to use of mini-Sdd7 in
348 *Agrobacterium*-mediated genome editing of rice plants and observed more mutants
349 recovered and a greater proportion of edited plants, which reflects both a higher
350 efficiency and lower toxicity compared to the most used hA3A-based CBE in
351 agricultural application (Figure S5G).

352 Soybean is one of the most important staple crops grown around the world,
353 serving as an essential source of vegetable oil and protein⁴⁷. Although previously
354 reported base editors have been widely used in many crops like rice, wheat, maize,
355 potato and more, cytosine base editing remains challenging and poorly efficient across

356 most sites tested in soybean crops^{32,48}. Since the first development of base editing,
357 only one article has used *Agrobacterium tumefaciens* to obtain stable transformations
358 and cytosine base-edited soybeans, but the efficiency was extremely low and resulted
359 in chimeric plants rather than completely edited soybeans³².

360 We wondered whether our newly developed Sdd-based CBEs would result in
361 superior cytosine base editing in soybeans. The transient base editing shown was
362 evaluated using a soybean hairy root transformation mediated by *Agrobacterium*
363 *rhizogenes*. This approach is often used in soybeans due to its quick nature (~20 days)
364 in allowing researchers to evaluate editing percentages in root cells. We constructed
365 vectors with an AtU6 promoter driving sgRNA expression and a CaMV 2 × 35S
366 promoter driving CBE expression and evaluated these using transgenic soybean hairy
367 roots following *Agrobacterium rhizogenes*-mediated transformations (Figure S5H).
368 We found that the APOBEC/AID deaminases had low editing activities across all five
369 sites evaluated as expected, including at the *GmALSI-T2* and *GmPPO2* sites which
370 were particularly difficult to edit by other CBEs in soybean (Figure 5D). Remarkably,
371 mini-Sdd7 displayed a 26.3-fold, 28.2-fold, and 10.8-fold increased cytosine base
372 editing levels, respectively, compared to rAPOBEC1, hA3A and human
373 activation-induced cytidine deaminase (hAID), respectively, across the five sites and
374 reaching editing efficiencies up to 67.4% (Figure 5D). However, the cells from hairy
375 root transformations are impossible to regenerate into soybean plants so the canonical
376 *Agrobacterium tumefaciens* is used to perform stable soybean plant editing in
377 cotyledons.

378 We next sought to use hA3A and mini-Sdd7 to base edit and obtain transgenic
379 soybean plants following *Agrobacterium tumefaciens*-mediated transformation. We
380 chose to edit the endogenous *GmPPO2* gene to create an R98C mutation, which
381 would result in carfentrazone-ethyl resistant soybean plants⁴⁹. Although the editing
382 efficiencies from hairy root transformations are a great approach for evaluating
383 relative editing efficiencies, it is not reflective of the percentage of edited plants
384 following soybean plant regeneration. Even with the highly efficient hA3A-base
385 editor in plants, we never successfully obtained cytosine base-edited plants (Figure

386 5E). Surprisingly, we obtained 34 base-edited heterozygotes from 154 transgenic
387 soybean seedlings of Sdd7 transgenic plants from four independent biological
388 experiments (Figure 5E). Therefore, Sdd7 now enables efficient cytosine base editing
389 in soybean plants, which will greatly contribute to future agricultural breeding efforts
390 (Figure 5E and 5F).

391 After treatment with carfentrazone-ethyl for ten days, we could obviously observe
392 that while the wild-type plant was sensitive to wilting and could not generate roots,
393 the mutated plant edited by Sdd7 grew well and normal (Figure 5G). The
394 development of efficient cytosine base editors for use in soybean plants could enable
395 diverse applications in the future.

396 Discussion

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

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

415 Currently only a few cytidine deaminases are in use as cytosine base editors.
416 Canonical efforts based solely on protein engineering and directed evolution have
417 helped diversify editing properties, however, these efforts are generally difficult to
418 establish. Using our structure-based clustering methods, we discovered and profiled a
419 suite of deaminases with distinct properties that can work both in plants and
420 mammalian cells.

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

435 Advances in sequencing methods have propelled the discovery of new species
436 and proteins. The advent of AI-assisted protein structure predictions in combination
437 with growing numbers of sequencing efforts will further spark new enzyme discovery
438 and enable even greater bioengineering efforts.

439 **Limitations of the study**

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

443 unknown protein families.

444 **Acknowledgement**

445 We thank Prof. Youwei Ai and Prof. Qingfeng Wu for kindly providing HEK293T and
446 N2a cell lines, respectively. We thank Prof. Qi-jun Chen for kindly providing
447 pBSE901. We thank Prof. Tianfu Han for kindly providing the seeds of
448 Zhonghuang13 soybean.

449 This work was supported by the National Natural Science Foundation of China
450 (32388201), the National Key Research and Development Program
451 (2022YFF1002802), the Ministry of Agriculture and Rural Affairs of China, and the
452 Strategic Priority Research Program of the Chinese Academy of Sciences (Precision
453 Seed Design and Breeding, XDA24020102). Q.L is supported by Postdoctoral
454 Innovative Talent Support Program of China (BX2021353) and China Postdoctoral
455 Science Foundation (2022M720163). K.T.Z. was supported by the Schmidt Science
456 Fellows.

457 **Author contributions**

458 J.Huang, K.T.Z. and C.G. conceived the project and designed the experiments.
459 J.Huang discovered the new deaminases. H.F. and Y.Li performed the structure-based
460 protein classification analysis. J.Huang, Q.L., and Z.H. performed the protoplasts
461 transformation and NGS data collection experiments. J.Huang, and Q.L. performed
462 the mammalian cell transfection and NGS data collection experiments. J.Huang, H.F.,
463 and Z.H. collected the TRAP-seq data. J.Huang and Q.G. analyzed the Novaseq and
464 Miseq data. G.L. and J.Hu prepared Miseq samples. H.F., and G.L. constructed the
465 binary vectors for rice and soybean plant transformation. B.L. obtained regenerated
466 rice plants. J.Huang, Z.H., and B.L. identified rice mutants. H.X., H.F., L.Z, Y.R.,
467 Z.H., and R.Z. performed soybean transformation, base-edited plants identification,
468 and soybean resistance experiments. Y.Luo, K.Q., and P.H. generated the HEK293T
469 cells with stable transfected TRAP-12K library. E.Z. provided AAV vector with guide
470 RNA for mouse targets. Q.L. and Z.H. prepared the figures. C.G. and K.T.Z. and

471 supervised the study. Q.L., H.F., Y.Li, K.T.Z. and C.G. wrote the manuscript with
472 input from all authors. J.-L.Q. revised the manuscript.

473 **Declaration of interests**

474 The authors have submitted two patent application based on the results reported in
475 this paper. K.T.Z. is a founder and employee at Qi Biodesign.

476 **Inclusion and diversity**

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

478 **Figure legends**

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

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

485 (B) Structural similarity matrix to reflect similarities between 242 predicted protein
486 (238 cytidine deaminases and 4 JAB) structures across 16 deaminase families and one
487 outgroup. Different family proteins are distinguished by different colors; heat map
488 color shades indicate the degree of similarity. (C) The classification of proteins into
489 different deaminase families based on protein structure and labeled with different
490 color modes.

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

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

493 (A) Classification of SCP1.201 deaminases based on protein structure. The JAB
494 families are colored brown and regarded as an outgroup, and the tested deaminases
495 are shown in red (single-strand editing), green (double-strand editing) or dark grey
496 (no editing). Undefined deaminases in light grey await further functional analysis.
497 (B) Predicted core structure of DddA by AlphaFold2.
498 (C) Characteristics of the canonical structure of Ddd protein.
499 (D) Predicted core structure of Sdd7 by AlphaFold2.
500 (E) Characteristics of the canonical structure of Sdd protein.
501 (F) Experimental evaluation of dsDNA deamination activity of Ddds at two
502 endogenous sites in HEK293T cells. The edited bases used for calculating editing are
503 highlighted in green.
504 (G) Experimental evaluation of ssDNA deamination activity of Sdds at two
505 endogenous sites in HEK293T cells. The edited bases used for calculating editing are
506 highlighted in green.
507 Data in (F) and (G) are representative of three independent biological replicates ($n =$
508 3).

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

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

513 (B) Plasmid library assay to profile context preferences of each Ddd protein in
514 mammalian cells. Candidate proteins target and edit the “NC₁₀N” motif.

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

517 For all plots, dots represent individual biological replicates, bars represent mean
518 values, and error bars represent the s.d. of three independent biological replicates ($n =$
519 3).

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

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

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

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

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

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

539 (F) On-target:off-target editing ratios of Sdd6, rAPOBEC1-YE1, rAPOBEC1-YEE,
540 rAPOBEC1, and hA3A tested across two on-target and three off-target sites in
541 HEK293T cells.

542 For (E) and (F), Dots represent individual biological replicates, bars represent mean
543 values, and error bars represent the s.d. of three independent biological replicates ($n =$
544 3). Data are presented as mean values \pm s.d. P values were obtained using two-sided
545 Mann-Whitney tests. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

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

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

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

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

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

560 (E) Frequencies of mutations induced by mini-Sdd7 and hA3A in T₀ stable soybean
561 plant editing in cotyledons by canonical *Agrobacterium tumefaciens*. The data were
562 collected by four independent biological experiments.

563 (F) The genotypes of base edited soybean plants.

564 (G) Phenotypes of soybean plants treated with carfentrazone-ethyl for 10-days. Left
565 panel, wild-type soybean plant (R98). Right panel, base-edited soybean plant (C98).
566 Bar=1 cm.

567 For (A), (C) and (D), Dots represent individual biological replicates, bars represent
568 mean values, and error bars represent the s.d. of three or four independent biological
569 replicates.

570 **STAR★Methods**

571 **RESOURCE AVAILABILITY**

572 **Lead contact**

573 Further information and requests for resources and reagents should be directed to and
574 will be fulfilled by the Lead Contact: Caixia Gao (cxgao@genetics.ac.cn).

575 **Materials availability**

576 All unique/stable reagents generated in this study are available from the Lead Contact
577 with a completed Materials Transfer Agreement.

578 **Data availability**

579 The deep amplicon sequencing data were deposited in the PRJNA915939,
580 PRJNA915940, PRJNA915941, and PRJNA915942. All other data are available in
581 the main paper or supplement.

582 **EXPERIMENTAL MODEL AND SUBJECT DETAILS**

583 ***E.coli* transfection**

584 FastT1 *E.coli* competent cells were used for amplifying plasmid DNA. Transfected
585 *E.coli* cells were grown at 37°C in Lysogeny Broth (LB) medium supplemented with
586 100 mg/mL ampicillin or kanamycin overnight.

587 **Rice protoplast transfection**

588 For protoplasts transfection, we used the Japonica rice (*Oryza sativa*) variety
589 Zhonghua11 to prepare protoplasts. Protoplast isolation and transformation were
590 performed as described previously⁵². Plasmids (5 µg per construct) were introduced
591 by PEG-mediated transfection. The transfected protoplasts were normally incubated at
592 26 °C for 72 hours for fluorescence cell observation or DNA extraction.

593 **Mammalian Cell lines and culture conditions**

594 Both human HEK293T cells (ATCC, CRL-3216) and mouse N2a cells (ATCC,
595 CCL-131) were cultured in Dulbecco's Modified Eagle's medium (DMEM, Gibco)
596 supplemented with 10% (vol/vol) fetal bovine serum (FBS, Gibco) and 1% (vol/vol)
597 Penicillin-Streptomycin (Gibco) in a humidified incubator at 37 °C with 5% CO₂.

598 **METHOD DETAILS**

599 **Protein clustering and analyzing**

600 Protein sequences were downloaded from InterPro database⁵³ and NCBI's BLAST⁵⁴
601 (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) on the NR database. HMM was utilized to
602 annotate deaminase domains to reduce the accumulation of unrelated information by
603 HMMER⁵⁵. We randomly chose 15 proteins from each family and clustered their
604 domain sequences with a threshold of 90% sequence identity and 90% coverage using
605 CD-HIT⁵⁶. Representatives of each cluster were selected for further analysis. High
606 confidence protein structures were predicted by Alphafold v2.2.0 and filtered with
607 average per-residue confidence metric pLDDT ≥ 70.

608 Multiple sequence alignment was performed using Multiple Protein Sequence
609 Alignment (MUSCLE)⁵⁷. The phylogenetic tree was constructed using IQ-TREE 2
610 (<http://www.iqtree.org>) with 1500 ultrafast bootstraps⁵⁸. A low perturbation strength
611 (-pers 0.2) and large number of stop iterations (-nstop 500) were set because of the
612 short length of the deaminase domains. Structure alignment was performed based on
613 normalized TM-score³³. The structural similarity matrix was further clustered by
614 Unweighted Pair Group Method with Arithmetic mean (UPGMA) and visualized by
615 Figtree (<http://tree.bio.ed.ac.uk/software/figtree/>). Protein structure diagrams were
616 made in PyMOL⁵⁹.

617 **Deaminase synthesis and removal of redundant sequence**

618 We chose gene fragments encoding complete deaminase domains as well as extra N
619 and C protein sequences for commercial synthesis (GenScript) (fig. S1). All of the
620 candidate cytidine deaminases were codon optimized (rice and wheat or human and
621 mouse). The toxin deaminase was split into two fragments and the split site was

622 selected according to DddA by protein structure alignment. The conserved protein
623 structure was obtained through multiple alignment of predicted structure in PyMOL⁵⁹,
624 which helps to conduct the removal of redundant sequence.

625 **Plasmid construction**

626 For plant CBE vectors (maize ubiquitin-1 promoter-driven CBEs), synthesized
627 deaminases were cloned into pnCas9-PBE vector (Addgene#98164), yielding vectors
628 with Ubi-1::NLS-deaminase-linker-nCas9(D10A)-UGI-NLS::CaMV expression
629 cassettes.

630 For CBE vectors for mammalian cells (CMV promoter-driven CBEs), synthesized
631 deaminases-SpCas9-2UGI were cloned into p2T-CMV-ABEmax-BlastR vector
632 (Addgene#152989), yielding vectors with
633 CMV::NLS-deaminase-linker-nCas9(D10A)-2xUGI-NLS::bGH expression cassettes.

634 The DdCBE vectors including NLS, TALE array sequences, candidate cytidine
635 deaminases, and UGI sequence were codon optimized for both human and mouse,
636 synthesized commercially (Genscript), and cloned into pCMV_BE4max vector
637 (Addgene#112093), yielding vectors with
638 CMV::NLS-TALE-deaminase-UGI-NLS::bGH expression cassettes.

639 The plant sgRNA vectors (rice U3 promoter drives sgRNA) were constructed as
640 reported previously using the pOsU3 backbone (Addgene#170132)⁶⁰. To construct
641 human and mouse sgRNA vectors (human U6 promoter drives sgRNA), the hU6
642 promoter was amplified and cloned into the pOsU3 backbone, followed by sgRNA
643 target sequence cloning steps⁵².

644 Plant SaCas9 vectors for off-target testing were constructed as reported
645 previously⁴².

646 To construct AAV vectors, the sequences between ITRs were synthesized (GenScript)
647 and cloned into pX601 vector (Addgene#61591), followed by sgRNA target sequence
648 cloning steps.

649 To construct binary vectors for rice plant transformation, the candidate cytidine
650 deaminases were codon optimized, synthesized commercially (GenScript), and cloned

651 into pH-nCas9-PBE vector (Addgene#98163), followed by sgRNA target sequence
652 cloning steps⁵².

653 To construct binary vectors for soybean hairy root transformation, NLS, candidate
654 cytidine deaminases, linker, nCas9(D10A), UGI, P2A, mScarlet sequences were
655 codon optimized, synthesized commercially (GenScript), and cloned into pBSE901
656 (Addgene#91709) vector, followed by sgRNA target sequence cloning steps. To
657 construct binary vectors for soybean transformation, the selection marker was
658 replaced by the *EPSPS* sequence.

659 **Mammalian cell line transfection**

660 All the cells were routinely tested for Mycoplasma contamination with a Mycoplasma
661 Detection Kit (Transgen Biotech). The cells were seeded into 48-well
662 Poly-D-Lysine-coated plates (Corning) in the absence of antibiotic. After 16-24 hours,
663 cells were incubated with 1 µL Lipofectamine 2000 (ThermoFisher Scientific), 300 ng
664 vector with deaminases, and 100 ng sgRNA expression vector. For DdCBEs
665 transfection, cells were incubated with 1 µL Lipofectamine 2000, 300ng TALE-L and
666 300ng TALE-R. 72 hours later the cells were washed with PBS, followed by DNA
667 extraction. For examining off-target effects by the R-loop assay, four vectors namely
668 BE4max vector, SaCas9BE4max vector and the corresponding sgRNA vectors were
669 co-transfected into cells³⁶.

670 **TRAPseq library**

671 We used the sgRNA 12K-TRAPseq library for evaluation of base editor properties.
672 We seeded 2×10^6 cells into 100 mm dish 20-hours before viral transduction. We
673 transduced 500 µL of sgRNA lentivirus. For stably integrated cells, we used 1 µg/mL
674 of puromycin (Gibco) to select. For each base editor, we seeded 2×10^6 cells into
675 6-plates 24-hours before transfection. We transfected 15 µg of each CBE member
676 plasmid DNA and 15 µg of Tol2 DNA with 60 µL of Lipofectamine 2000. Following
677 24 hours after transfection, we changed new culturing media to contain 10 µg/mL
678 blasticidin (Gibco). After another 3 days, we washed the cells, suspended and
679 reseeded all cells in 10 µg/mL blasticidin-containing media. After 6 days, we

680 harvested all cells by washing with PBS then centrifuged and extracted DNA using
681 Cell/Tissue DNA Isolation Mini Kit (Vazyme). For each member, we prepared
682 sequencing reactions by applying 1.2 µg of DNA with a first set of primers following
683 by barcoding and next-generating sequencing.

684 **DNA extraction**

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

690 **Amplicon deep sequencing and data analysis**

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

694 Nested PCR was used for amplification. In the first round PCR, the target region
695 was amplified from genomic DNA with site-specific primers. In the second round,
696 both forward and reverse barcodes were added to the ends of the PCR products for
697 library construction. Equal amounts of PCR product were pooled and purified with a
698 GeneJET Gel Extraction Kit (Thermo Scientific) and quantified with a NanoDrop
699 2000 spectrophotometer (Thermo Scientific). The purified products were sequenced
700 commercially using the NovaSeq or Miseq platform, and the sequences around the
701 target regions were examined for editing events⁶⁰. Amplicon sequencing was repeated
702 three times for each target site using genomic DNA extracted from three independent
703 samples. Analysis of base editing behaviour by NovaSeq and Miseq was performed as
704 described previously⁶⁰.

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

709 ***Agrobacterium*-mediated transformation of rice calli**

710 The Japonica rice (*Oryza sativa*) variety Zhonghua 11 was used for genetic
711 transformation in this study. Binary vectors were introduced into *Agrobacterium*
712 *tumefaciens* strain AGL1 by electroporation. *Agrobacterium*-mediated transformation
713 of Zhonghua11 callus cells was conducted as reported⁶¹. Hygromycin (50 µg/ml) was
714 used to select transgenic plants.

715 **Soybean hairy root transformation and plant transformation**

716 The soybean (*Glycine max*) variety Williams 82 was used to generate hairy roots.
717 Binary vectors were introduced into *Agrobacterium rhizogenes* strain K599 by
718 electroporation. Explants were allowed to grow and develop roots for around 20 days
719 in germination medium. Transgenic hairy roots were generated without selection in
720 10-12 days⁶². The soybean (*Glycine max*) variety Zhonghuang13 were used for
721 generation of transgenic plants using *Agrobacterium tumefaciens*-mediated stable
722 transformation. 10 mg/L glyphosate was used for selection during plant regeneration⁶³.
723 For phenotype identification of base-edited soybean, 0.3 mg/L carfentrazone-ethyl
724 were added in rooting medium for selection.

725 **Plant mutant identification**

726 Genomic DNA of transgenic plants was extracted with DNA Quick Plant System
727 (Tiangen Biotech). Specific primers were used to amplify and sequence the target
728 sites as described previously⁶⁰ (Supplementary Table 1) (BGI). T₀ transgenic rice and
729 soybean plants were examined individually.

730 **Statistical analysis**

731 All numerical values are presented as means ± s.d. Significant differences between
732 controls and treatments were tested using the two-sided Mann-Whitney test, and $P <$
733 0.05 was considered statistically significant, $P < 0.01$ was considered statistically
734 extremely significant.

735 **Supplemental information**

736 **Primary Supplemental PDF**

737 **Supplementary Table 1.** Primers used in this study.

738 **References**

- 739 1. Sharifi, F., and Ye, Y. (2022). Identification and classification of reverse
740 transcriptases in bacterial genomes and metagenomes. *Nucleic Acids Res.* **50**, e29.
741 10.1093/nar/gkab1207.
- 742 2. Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns,
743 S. J. J., Charpentier, E., Cheng, D., Haft, D. H., Horvath, P., et al. (2020).
744 Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived
745 variants. *Nat. Rev. Microbiol.* **18**, 67-83. 10.1038/s41579-019-0299-x.
- 746 3. Berntsson, R.P., Smits, S.H., Schmitt, L., Slotboom, D.J., and Poolman, B. (2010).
747 A structural classification of substrate-binding proteins. *FEBS Lett.* **584**,
748 2606-2617. 10.1016/j.febslet.2010.04.043.
- 749 4. Chandonia, J.M., Guan L., Lin S., Yu C., Fox N.K., Brenner S.E., et al. (2022).
750 SCOPe: improvements to the structural classification of proteins - extended
751 database to facilitate variant interpretation and machine learning. *Nucleic Acids
752 Res.* **50**, D553-D559. 10.1093/nar/gkab1054.
- 753 5. wwPDB consortium, (2019). Protein Data Bank: the single global archive for 3D
754 macromolecular structure data. *Nucleic Acids Res.* **47**, D520-D528.
- 755 6. Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M. L., Burdett, T.,
756 Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L. J., et al. (2023). MGnify:
757 the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **47**,
758 D520-D528. 10.1093/nar/gkac1080.
- 759 7. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,
760 Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly
761 accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589.
762 10.1038/s41586-021-03819-2.
- 763 8. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R.,
764 Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., et al. (2021). Accurate
765 prediction of protein structures and interactions using a three-track neural network.
766 *Science* **373**, 871-876. 10.1126/science.abj8754.
- 767 9. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G.,

- 768 Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein
769 Structure Database: massively expanding the structural coverage of
770 protein-sequence space with high-accuracy models. *Nucleic Acids Res.* *50*,
771 D439-D444. 10.1093/nar/gkab1061.
- 772 10. Mok, B. Y., de Moraes, M. H., Zeng, J., Bosch, D. E., Kotrys, A. V., Raguram, A.,
773 Hsu, F., Radey, M. C., Peterson, S. B., Mootha, V. K., et al. (2020). A bacterial
774 cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature*
775 *583*, 631-637. 10.1038/s41586-020-2477-4.
- 776 11. Zhang, H., Yang, B., Pomerantz, R. J., Zhang, C., Arunachalam, S. C., and Gao, L.
777 (2003). The cytidine deaminase CEM15 induces hypermutation in newly
778 synthesized HIV-1 DNA. *Nature* *424*, 94-98. 10.1038/nature01707.
- 779 12. Weiss, B. (2007). The deoxycytidine pathway for thymidylate synthesis in
780 *Escherichia coli*. *J. Bacteriol.* *189*, 7922-7926. 10.1128/JB.00461-07.
- 781 13. Esnault, C., Heidmann, O., Delebecque, F., Dewannieux, M., Ribet, D., Hance, A.
782 J., Heidmann, T., and Schwartz, O. (2005). APOBEC3G cytidine deaminase
783 inhibits retrotransposition of endogenous retroviruses. *Nature* *433*, 430-433.
784 10.1038/nature03238.
- 785 14. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016).
786 Programmable editing of a target base in genomic DNA without double-stranded
787 DNA cleavage. *Nature* *533*, 420-424. 10.1038/nature17946.
- 788 15. Nishida, K., Arazoe, T., Yachie, N., Banno, S., Kakimoto, M., Tabata, M.,
789 Mochizuki, M., Miyabe, A., Araki, M., Hara, K. Y., et al. (2016). Targeted
790 nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune
791 systems. *Science* *353*, aaf8729. 10.1126/science.aaf8729.
- 792 16. Cox, D. B. T., Gootenberg, J. S., Abudayyeh, O. O., Franklin, B., Kellner, M. J.,
793 Joung, J., and Zhang, F., et al. (2017). RNA editing with CRISPR-Cas13. *Science*
794 *358*, 1019-1027. 10.1126/science.aaq0180.
- 795 17. Harris, R.S., Petersen-Mahrt, S.K., and Neuberger, M.S. (2002). RNA editing
796 enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol. Cell*
797 *10*, 1247-1253. 10.1016/s1097-2765(02)00742-6.
- 798 18. Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., Liu, K. I.,
799 Zhang, R., Ramaswami, G., Ariyoshi, K., et al. (2017). Dynamic landscape and
800 regulation of RNA editing in mammals. *Nature* *550*, 249-254.
801 10.1038/nature24041.
- 802 19. Wolf, J., Gerber, A. P., and Keller, W. (2002). *tadA*, an essential tRNA-specific
803 adenosine deaminase from *Escherichia coli*. *EMBO J.* *21*, 3841-3851.
804 10.1093/emboj/cdf362.
- 805 20. Iyer, L.M., Zhang, D., Rogozin, I.B., and Aravind, L. (2011). Evolution of the
806 deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic
807 acid deaminases from bacterial toxin systems. *Nucleic Acids Res.* *39*, 9473-9497.

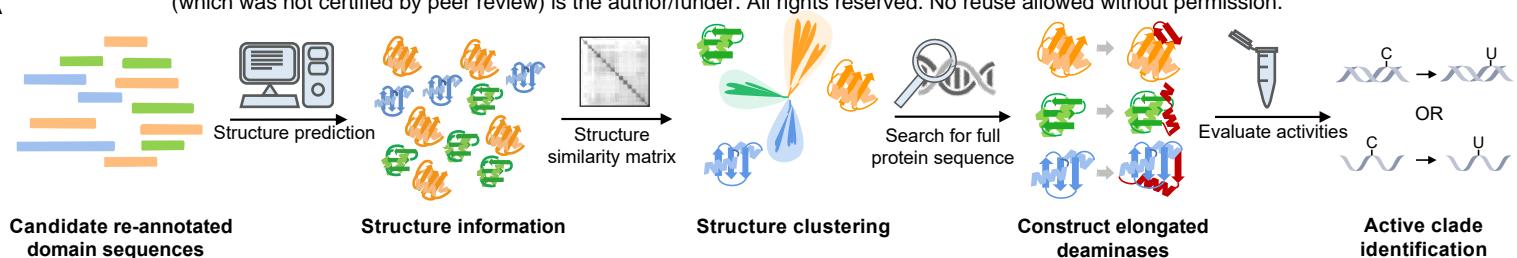
- 808 10.1093/nar/gkr691.
- 809 21. Krishnan, A., Iyer, L.M., Holland, S.J., Boehm, T., and Aravind, L. (2018).
810 Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of
811 clades and widespread roles in immunity. *Proc. Natl. Acad. Sci. U S A* *115*,
812 E3201-E3210. 10.1073/pnas.1720897115.
- 813 22. Gao, C. (2021). Genome engineering for crop improvement and future agriculture.
814 *Cell* *184*, 1621-1635. 10.1016/j.cell.2021.01.005.
- 815 23. Anzalone, A.V., Koblan, L.W., and Liu, D.R. (2020). Genome editing with
816 CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat.*
817 *Biotechnol.* *38*, 824-844. 10.1038/s41587-020-0561-9.
- 818 24. Li, Y., Li, W., and Li, J. (2021). The CRISPR/Cas9 revolution continues: From
819 base editing to prime editing in plant science. *J. Genet. Genomics* *48*, 661-670.
820 10.1016/j.jgg.2021.05.001.
- 821 25. Zhang, R., Chen, S., Meng, X., Chai, Z., Wang, D., Yuan, Y., Chen, K., Jiang, L.,
822 Li, J., and Gao, C. (2021). Generating broad-spectrum tolerance to ALS-inhibiting
823 herbicides in rice by base editing. *Sci. China Life Sci.* *64*, 1624-1633.
824 10.1007/s11427-020-1800-5.
- 825 26. Chen, Y., Wang, Z., Ni, H., Xu, Y., Chen, Q., and Jiang, L. (2017).
826 CRISPR/Cas9-mediated base-editing system efficiently generates gain-of-function
827 mutations in *Arabidopsis*. *Sci. China Life Sci.* *60*, 520-523.
828 10.1007/s11427-017-9021-5.
- 829 27. Ma, Y., Zhang, J., Yin, W., Zhang, Z., Song, Y., and Chang, X. (2016). Targeted
830 AID-mediated mutagenesis (TAM) enables efficient genomic diversification in
831 mammalian cells. *Nat. Methods* *13*, 1029-1035. 10.1038/nmeth.4027
- 832 28. Hess, G. T., Frésard, L., Han, K., Lee, C. H., Li, A., Cimprich, K. A.,
833 Montgomery, S. B., and Bassik, M. C. (2016). Directed evolution using
834 dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* *13*,
835 1036-1042. 10.1038/nmeth.4038.
- 836 29. Yu, Y., Leete, T. C., Born, D. A., Young, L., Barrera, L. A., Lee, S. J., Rees, H. A.,
837 Ciaramella, G., and Gaudelli, N. M. (2020). Cytosine base editors with minimized
838 unguided DNA and RNA off-target events and high on-target activity. *Nat.*
839 *Commun.* *11*, 2052. 10.1038/s41467-020-15887-5.
- 840 30. Cheng, T. L., Li, S., Yuan, B., Wang, X., Zhou, W., and Qiu, Z. (2019).
841 Expanding C-T base editing toolkit with diversified cytidine deaminases. *Nat.*
842 *Commun.* *10*, 3612. 10.1038/s41467-019-11562-6.
- 843 31. Levy, J. M., Yeh, W. H., Pendse, N., Davis, J. R., Hennessey, E., Butcher, R.,
844 Koblan, L. W., Comander, J., Liu, Q., and Liu, D. R. (2020). Cytosine and adenine
845 base editing of the brain, liver, retina, heart and skeletal muscle of mice via
846 adeno-associated viruses. *Nat. Biomed. Eng.* *4*, 97-110.
847 10.1038/s41551-019-0501-5.

- 848 32. Cai, Y., Chen, L., Zhang, Y., Yuan, S., Su, Q., Sun, S., Wu, C., Yao, W., Han, T.,
849 and Hou, W. (2020). Target base editing in soybean using a modified
850 CRISPR/Cas9 system. *Plant Biotechnol. J.* *18*, 1996-1998. 10.1111/pbi.13386.
- 851 33. Zhang, C. Shine, M. Pyle, A.M. and Zhang, Y. (2022). US-align: universal
852 structure alignments of proteins, nucleic acids, and macromolecular complexes.
853 *Nat. Methods* *19*, 1109-1115. 10.1038/s41592-022-01585-1.
- 854 34. Zong, Y., Wang, Y., Li, C., Zhang, R., Chen, K., Ran, Y., Qiu, J. L., Wang, D., and
855 Gao, C. (2017). Precise base editing in rice, wheat and maize with a Cas9-cytidine
856 deaminase fusion. *Nat. Biotechnol.* *35*, 438-440. 10.1038/nbt.3811.
- 857 35. Mok, B. Y., Kotrys, A. V., Raguram, A., Huang, T. P., Mootha, V. K., and Liu, D.
858 R. (2022). CRISPR-free base editors with enhanced activity and expanded
859 targeting scope in mitochondrial and nuclear DNA. *Nat. Biotechnol.* *40*,
860 1378-1387. 10.1038/s41587-022-01256-8.
- 861 36. Koblan, L. W., Doman, J. L., Wilson, C., Levy, J. M., Tay, T., Newby, G. A.,
862 Maianti, J. P., Raguram, A., and Liu, D. R. (2018). Improving cytidine and
863 adenine base editors by expression optimization and ancestral reconstruction. *Nat.*
864 *Biotechnol.* *36*, 843-846. 10.1038/nbt.4172.
- 865 37. Zong, Y., Song, Q., Li, C., Jin, S., Zhang, D., Wang, Y., Qiu, J. L., and Gao, C.
866 (2018). Efficient C-to-T base editing in plants using a fusion of nCas9 and human
867 APOBEC3A. *Nat. Biotechnol.* *36*, 950-953. 10.1038/nbt.4261.
- 868 38. Lin, Q., Zhu, Z., Liu, G., Sun, C., Lin, D., Xue, C., Li, S., Zhang, D., Gao, C.,
869 Wang, Y., et al. (2021). Genome editing in plants with MAD7 nuclease. *J. Genet.*
870 *Genomics* *48*, 444-451. 10.1016/j.jgg.2021.04.003.
- 871 39. Xiang, X., Qu, K., Liang, X., Pan, X., Wang, J., Han, P., Dong, Z., Liu, L., Zhong,
872 J., Ma, T., Wang, Y., et al. (2020). Massively parallel quantification of CRISPR
873 editing in cells by TRAP-seq enables better design of Cas9, ABE, CBE gRNAs of
874 high efficiency and accuracy. *bioRxiv*, <https://doi.org/10.1101/2020.05.20.103614>.
- 875 40. Jin, S., Zong, Y., Gao, Q., Zhu, Z., Wang, Y., Qin, P., Liang, C., Wang, D., Qiu, J.
876 L., Zhang, F., et al. (2019). Cytosine, but not adenine, base editors induce
877 genome-wide off-target mutations in rice. *Science* *364*, 292-295.
878 10.1126/science.aaw7166.
- 879 41. Zuo, E., Sun, Y., Wei, W., Yuan, T., Ying, W., Sun, H., Yuan, L., Steinmetz, L. M.,
880 Li, Y., and Yang, H. (2019). Cytosine base editor generates substantial off-target
881 single-nucleotide variants in mouse embryos. *Science* *364*, 289-292.
882 10.1126/science.aav9973.
- 883 42. Jin, S., Fei, H., Zhu, Z., Luo, Y., Liu, J., Gao, S., Zhang, F., Chen, Y. H., Wang, Y.,
884 and Gao, C. (2020). Rationally designed APOBEC3B cytosine base editors with
885 improved specificity. *Mol. Cell* *79*, 728-740.e6. 10.1016/j.molcel.2020.07.005.
- 886 43. Doman, J.L., Raguram, A., Newby, G.A., and Liu, D.R. (2020). Evaluation and
887 minimization of Cas9-independent off-target DNA editing by cytosine base editors.

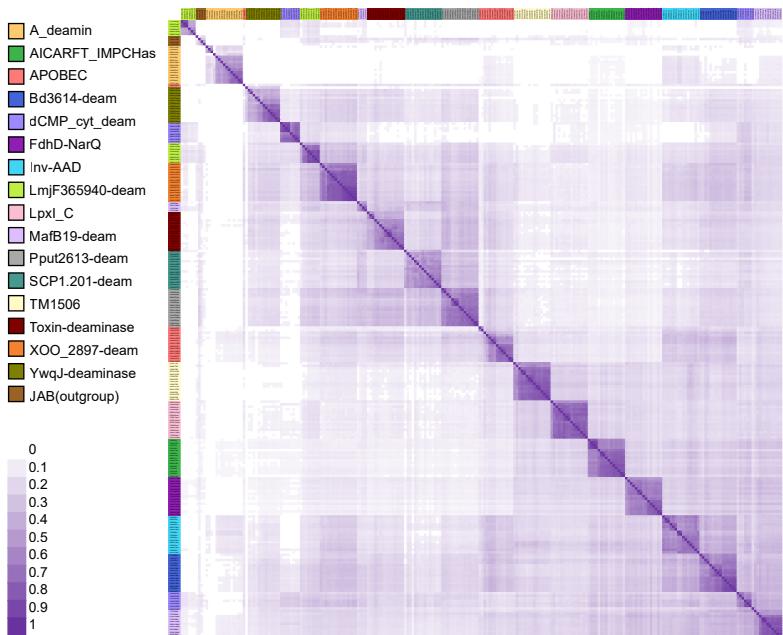
- 888 Nat. Biotechnol. 38, 620-628. 10.1038/s41587-020-0414-6.
- 889 44. Davis, J. R., Wang, X., Witte, I. P., Huang, T. P., Levy, J. M., Raguram, A.,
890 Banskota, S., Seidah, N. G., Musunuru, K., and Liu, D. R. (2022). Efficient in
891 vivo base editing via single adeno-associated viruses with size-optimized genomes
892 encoding compact adenine base editors. Nat. Biomed. Eng. 6, 1272-1283.
893 10.1038/s41551-022-00911-4.
- 894 45. Li, A., Mitsunobu, H., Yoshioka, S., Suzuki, T., Kondo, A., and Nishida, K.
895 (2022). Cytosine base editing systems with minimized off-target effect and
896 molecular size. Nat. Commun. 13, 4531. 10.1038/s41467-022-32157-8.
- 897 46. Pankowicz, F. P., Barzi, M., Legras, X., Hubert, L., Mi, T., Tomolonis, J. A.,
898 Ravishankar, M., Sun, Q., Yang, D., Borowiak, M., et al. (2016). Reprogramming
899 metabolic pathways in vivo with CRISPR/Cas9 genome editing to treat hereditary
900 tyrosinaemia. Nat. Commun. 7, 12642. 10.1038/ncomms12642.
- 901 47. Liu, S., Zhang, M., Feng, F., and Tian, Z. (2020). Toward a "Green Revolution"
902 for soybean. Mol. Plant 13, 688-697. 10.1016/j.molp.2020.03.002.
- 903 48. Molla, K.A., Sretenovic, S., Bansal, K.C. & Qi, Y. Precise plant genome editing
904 using base editors and prime editors. Nat. Plants 631 7, 1166-1187 (2021).
905 10.1038/s41477-021-00991-1
- 906 49. Dayan, F.E., Barker, A., and Tranel, P.J. (2018). Origins and structure of
907 chloroplastic and mitochondrial plant protoporphyrinogen oxidases: implications
908 for the evolution of herbicide resistance. Pest Manag. Sci. 74, 2226-2234.
909 10.1002/ps.4744.
- 910 50. Thompson, M.C., Yeates, T.O., and Rodriguez, J.A. (2020). Advances in methods
911 for atomic resolution macromolecular structure determination. F1000Res. 9, 667.
912 10.12688/f1000research.25097.1.
- 913 51. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C.,
914 Žídek, A., Nelson, A. W. R., Bridgland, A., et al. (2020). Improved protein
915 structure prediction using potentials from deep learning. Nature 577, 706-710.
916 10.1038/s41586-019-1923-7.
- 917 52. Shan, Q., Wang, Y., Li, J., and Gao, C. (2014). Genome editing in rice and wheat
918 using the CRISPR/Cas system. Nat. Protoc. 9, 2395–2410.
919 10.1038/nprot.2014.157.
- 920 53. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D.,
921 Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. (2009). InterPro: the
922 integrative protein signature database. Nucleic Acids Res. 37, D211–D215.
923 10.1093/nar/gkn785.
- 924 54. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and
925 Madden, T. L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res.
926 36, W5–W9. 10.1093/nar/gkn201.
- 927 55. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive
928 sequence similarity searching. Nucleic Acids Res. 39, W29–W37.
929 10.1093/nar/gkr367.

- 930 56. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for
931 clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152.
932 10.1093/bioinformatics/bts565.
- 933 57. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy
934 and high throughput. *Nucleic Acids Res.* 32, 1792–1797. 10.1093/nar/gkh340.
- 935 58. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D.,
936 von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient
937 methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37,
938 1530–1534. 10.1093/molbev/msaa015.
- 939 59. DeLano, W.L. (2000). The PyMOL molecular graphics system. Schrödinger LLC,
940 New York, NY, USA.
- 941 60. Jin, S., Lin, Q., Gao, Q., and Gao, C. (2022). Optimized prime editing in monocot
942 plants using PlantPegDesigner and engineered plant prime editors (ePPEs). *Nat.*
943 *Protoc.* 10.1038/s41596-022-00773-9.
- 944 61. Jin, S., Gao, Q., and Gao, C. (2021). An unbiased method for evaluating the
945 genome-wide specificity of base editors in rice. *Nat. Protoc.* 16, 431–457.
946 10.1038/s41596-020-00423-y.
- 947 62. Li, C., Zhang, H., Wang, X., and Liao, H. (2014). A comparison study of
948 *Agrobacterium*-mediated transformation methods for root-specific promoter
949 analysis in soybean. *Plant Cell Rep.* 33, 1921–1932. 10.1007/s00299-014-1669-5.
- 950 63. Li, S., Cong, Y., Liu, Y., Wang, T., Shuai, Q., Chen, N., Gai, J., and Li, Y. (2017).
951 Optimization of Agrobacterium-Mediated Transformation in Soybean. *Front. Plant*
952 *Sci.* 8, 246. 10.3389/fpls.2017.00246.

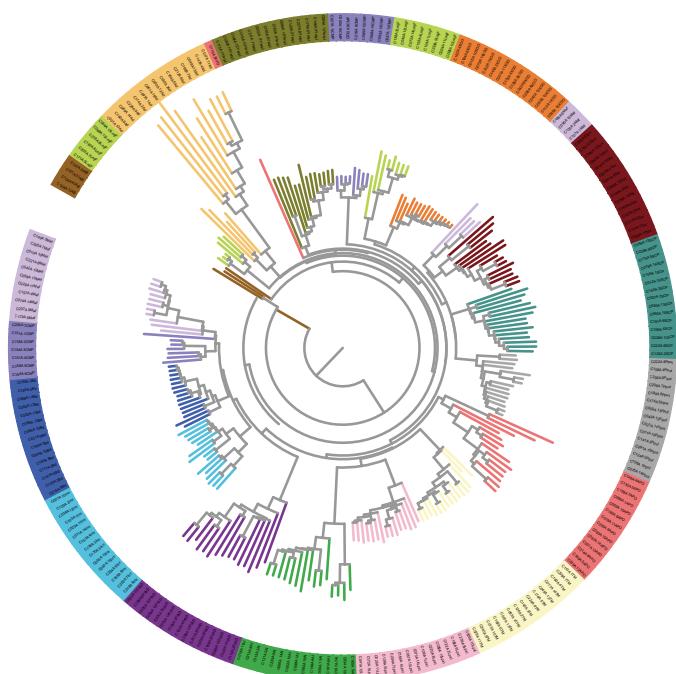
A



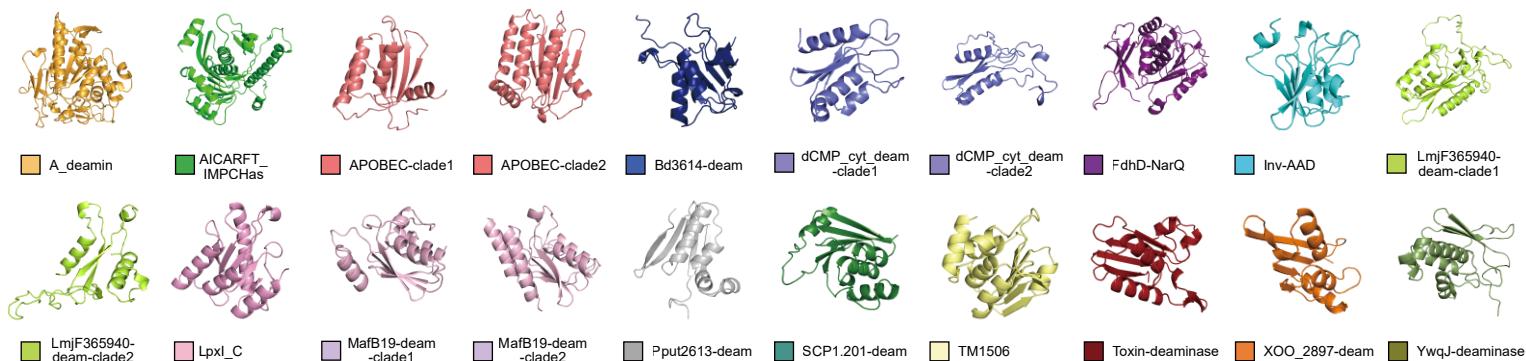
B



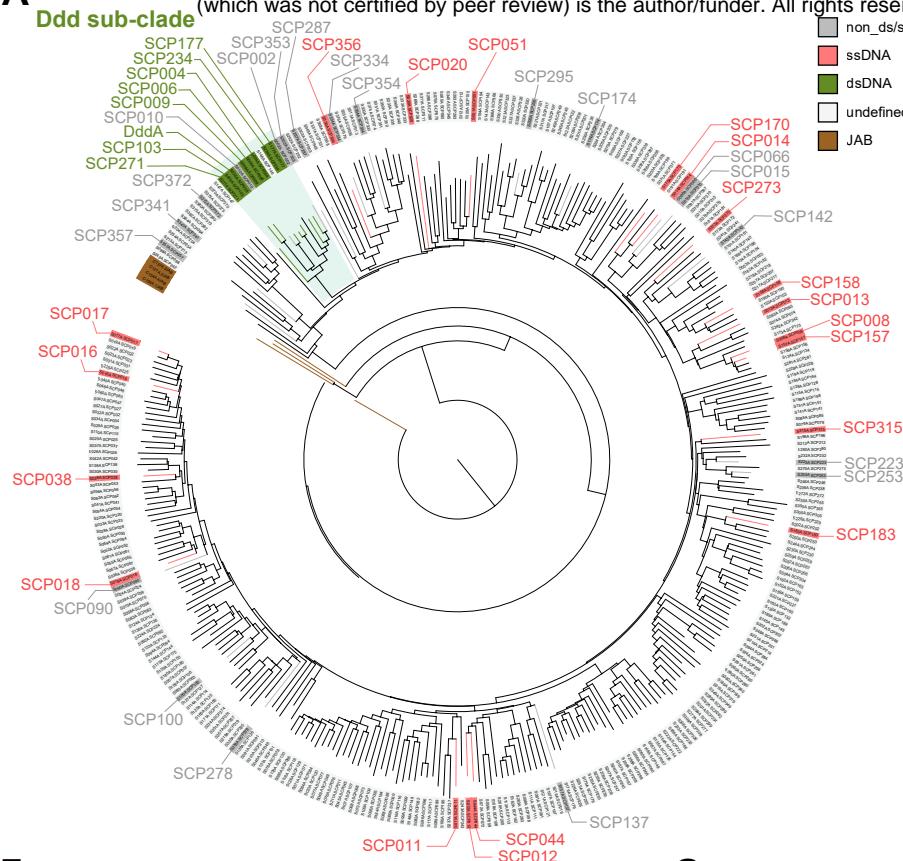
C



D



A



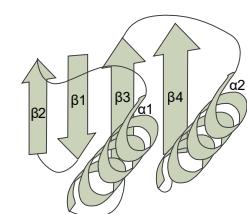
B

Predicted core structure of DddA



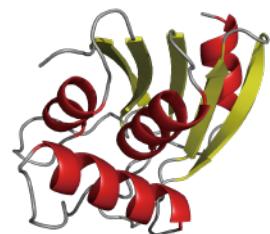
C

Canonical structure of Ddd deaminase



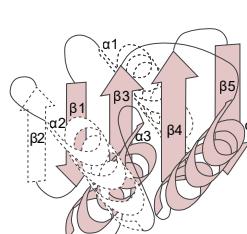
D

Predicted core structure of Sdd7

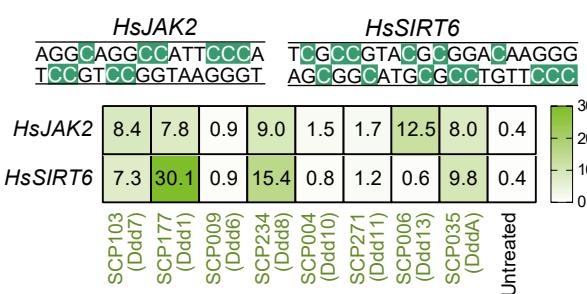


E

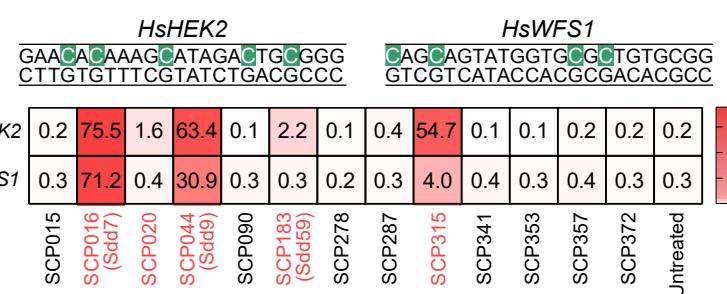
Canonical structure of Sdd deaminase



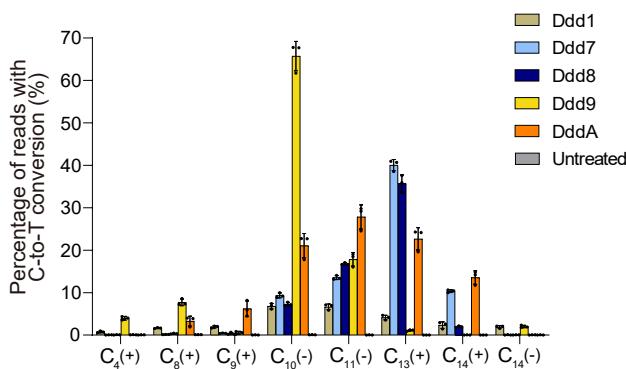
F



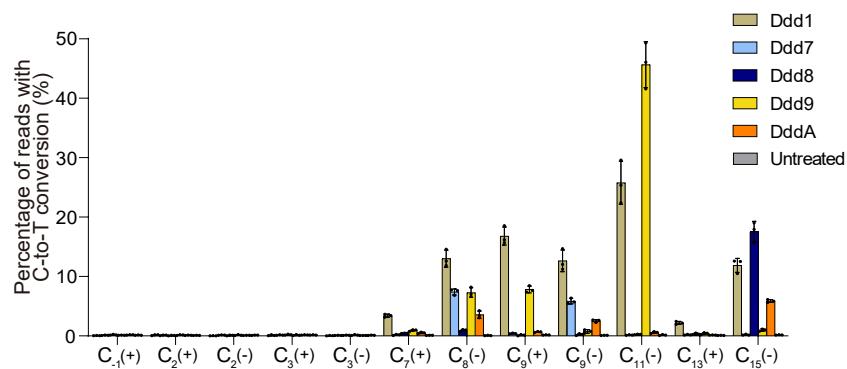
G



AG G C A G G C₈C₉ATT C₁₃C₁₄CA
TC C₁₄G T C₁₁C₁₀GG TAA G G GT



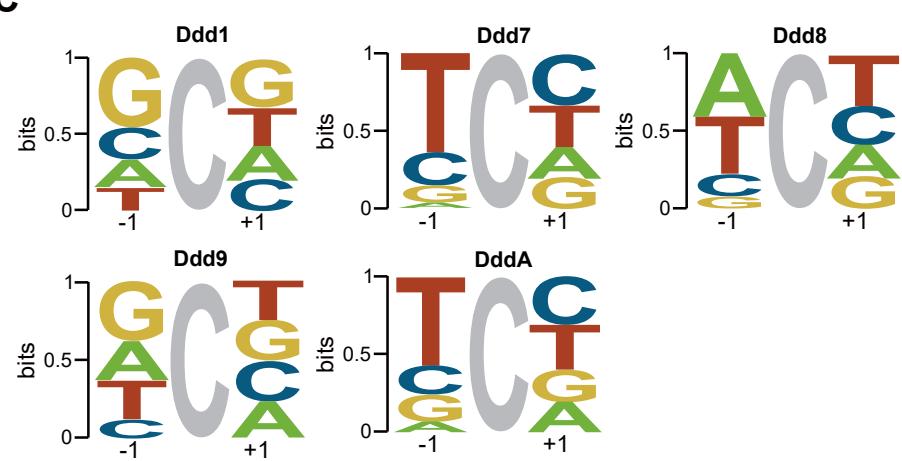
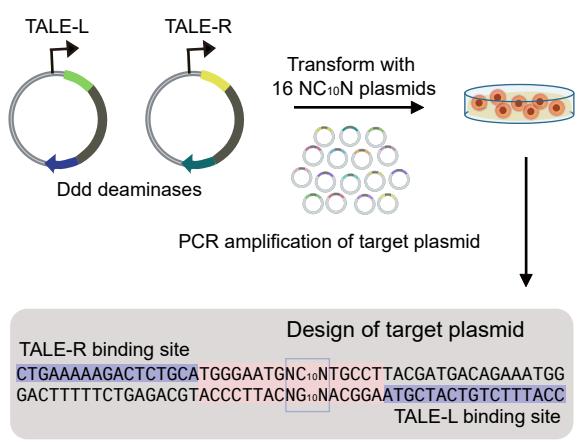
T C₋₁ G C₂ C₃ G T A C₇ G C₉ G G A C₁₃ AA G G G
A G C G G C₁₅ AT G C₁₁ G C₉ C₈ T G TT C₃ C₂ C

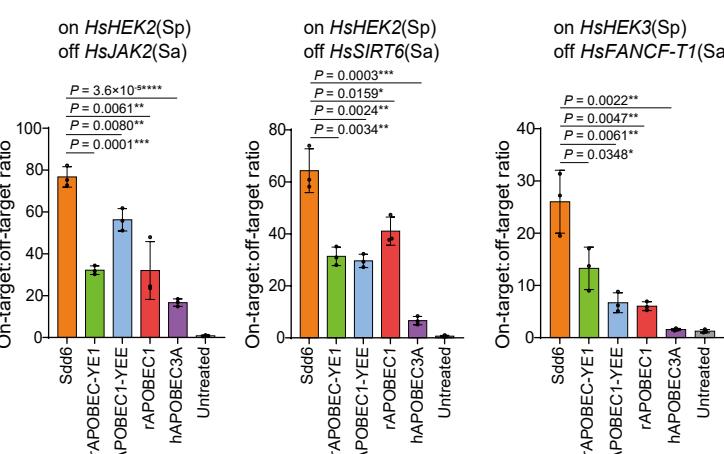
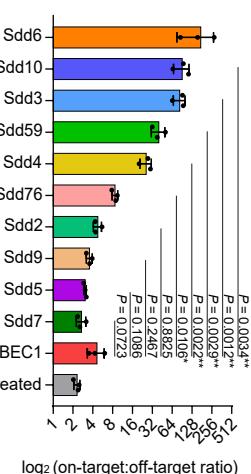
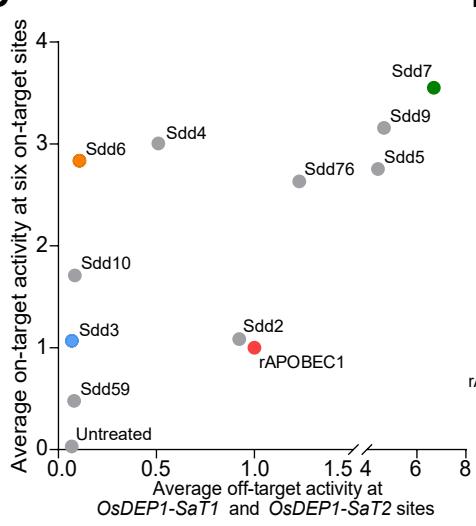
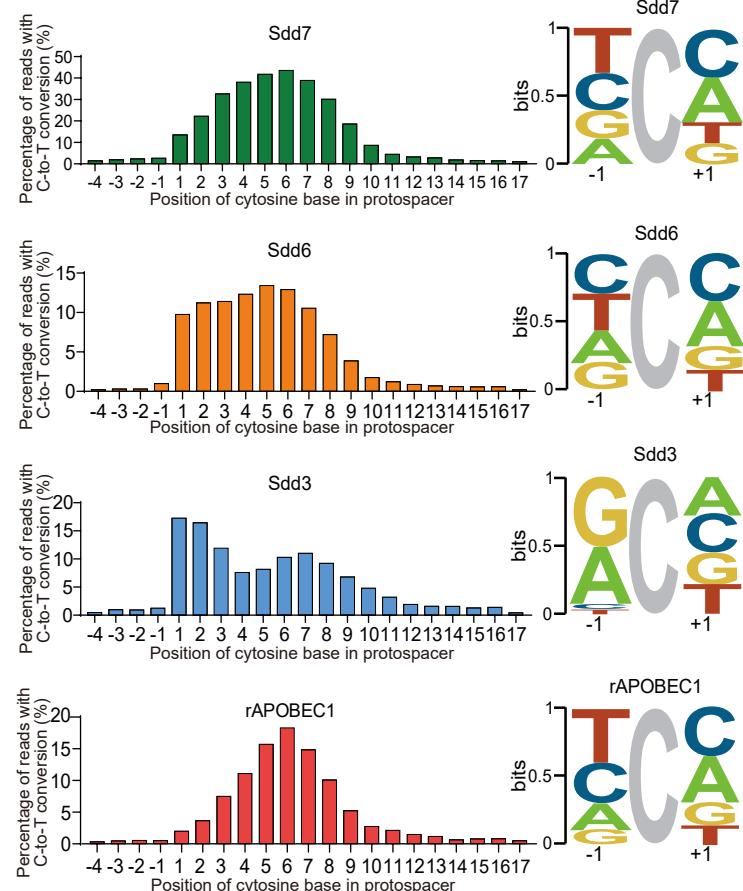
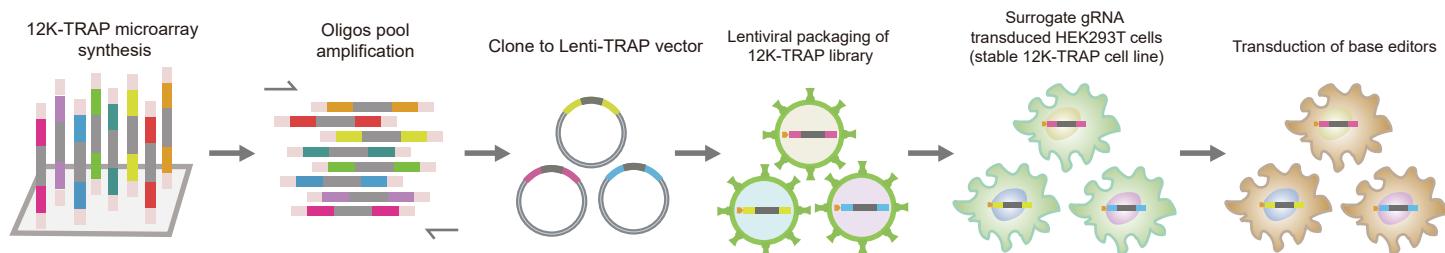
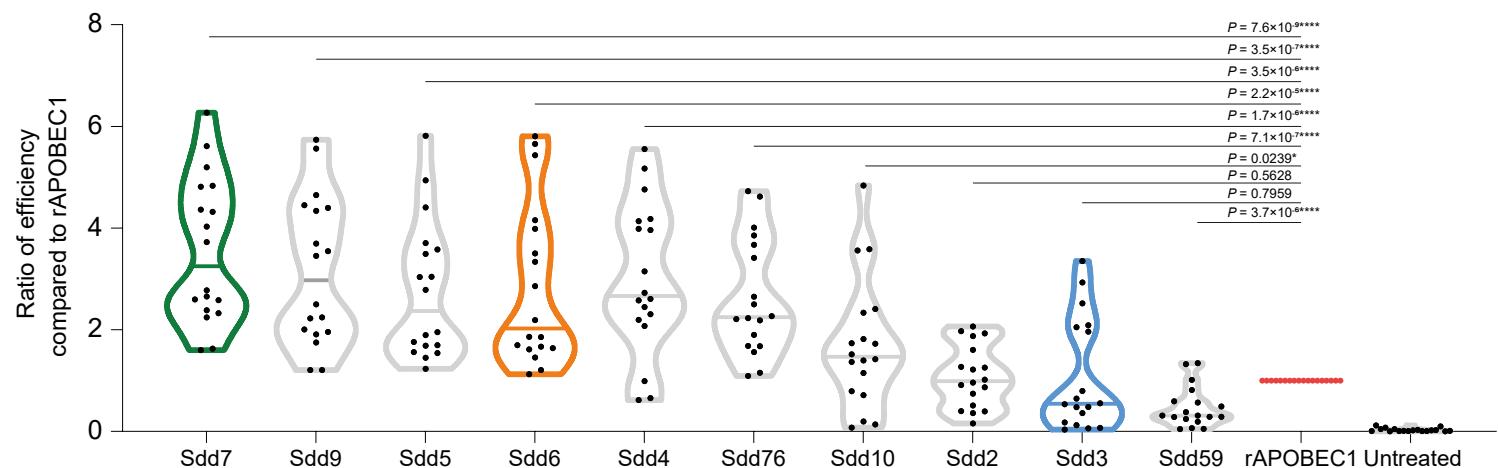


A

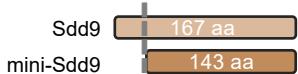
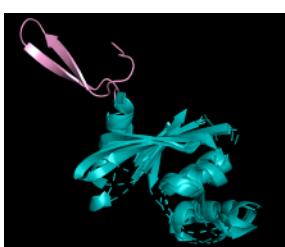
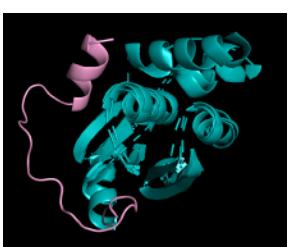
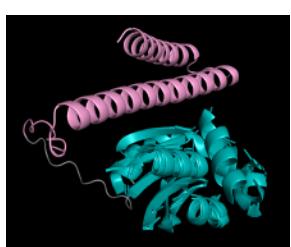
B

C

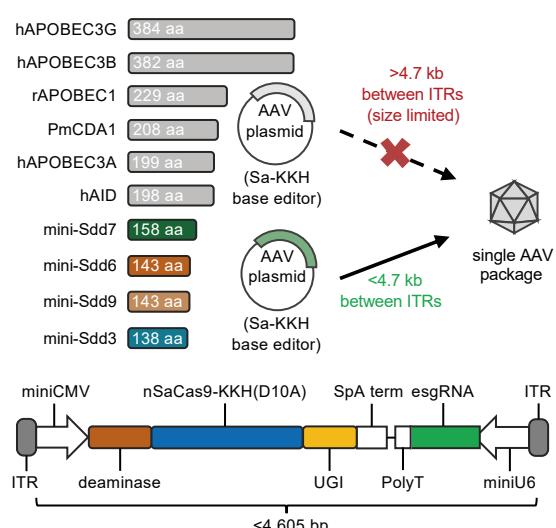




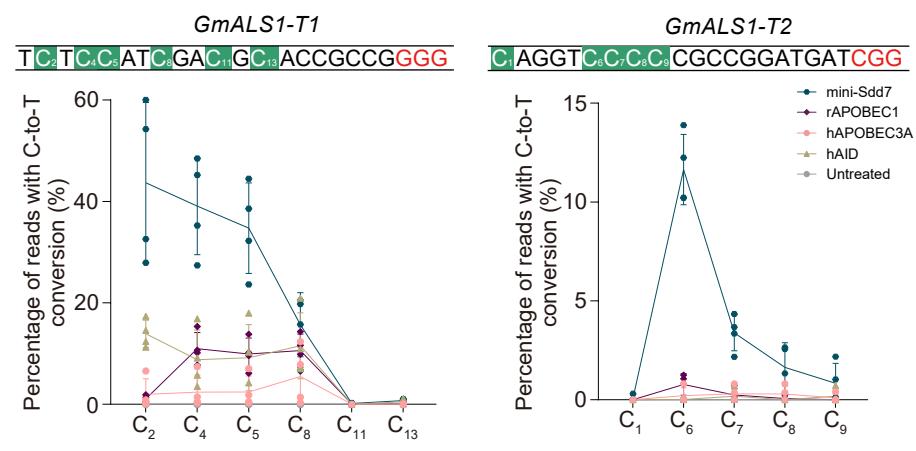
A



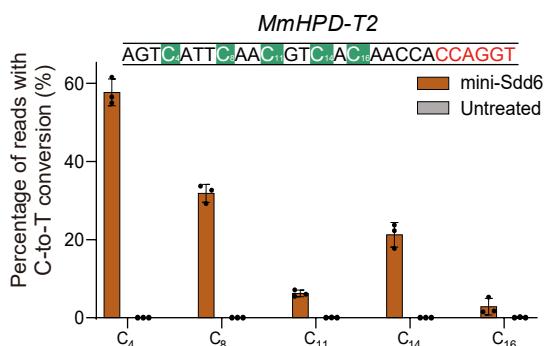
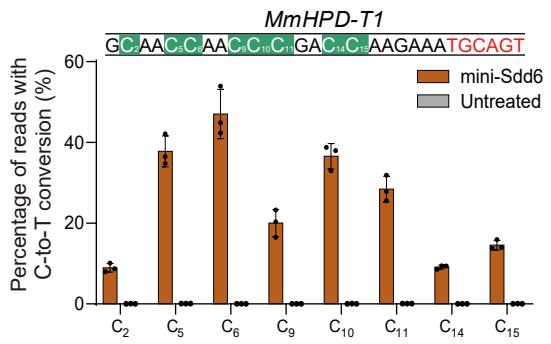
B



D



C



E

Repeat	Deaminase	Target gene	No. of transgenic plants	No. of mutants	Editing efficiency
Exp. 1	mini-Sdd7	GmPPO2	40	7	17.5%
Exp. 2	mini-Sdd7	GmPPO2	31	0	0.0%
Exp. 3	mini-Sdd7	GmPPO2	30	11	36.7%
Exp. 4	mini-Sdd7	GmPPO2	45	0	0.0%
	APOBEC3A	GmPPO2	7	3	42.9%
	APOBEC3A	GmPPO2	8	0	0.0%
	mini-Sdd7	GmPPO2	77	13	16.9%
	APOBEC3A	GmPPO2	61	0	0.0%

F



G

