nature communications

Article

https://doi.org/10.1038/s41467-025-57154-5

Improving adenine base editing precision by enlarging the recognition domain of CRISPR-Cas9

Received: 2 June 2024

Accepted: 11 February 2025

Published online: 28 February 2025

Check for updates

Shuliang Gao, Benson Weng, Douglas Wich, Liam Power, Mengting Chen, Huiwen Guan, Zhongfeng Ye, Chutian Xu & Qiaobing Xu

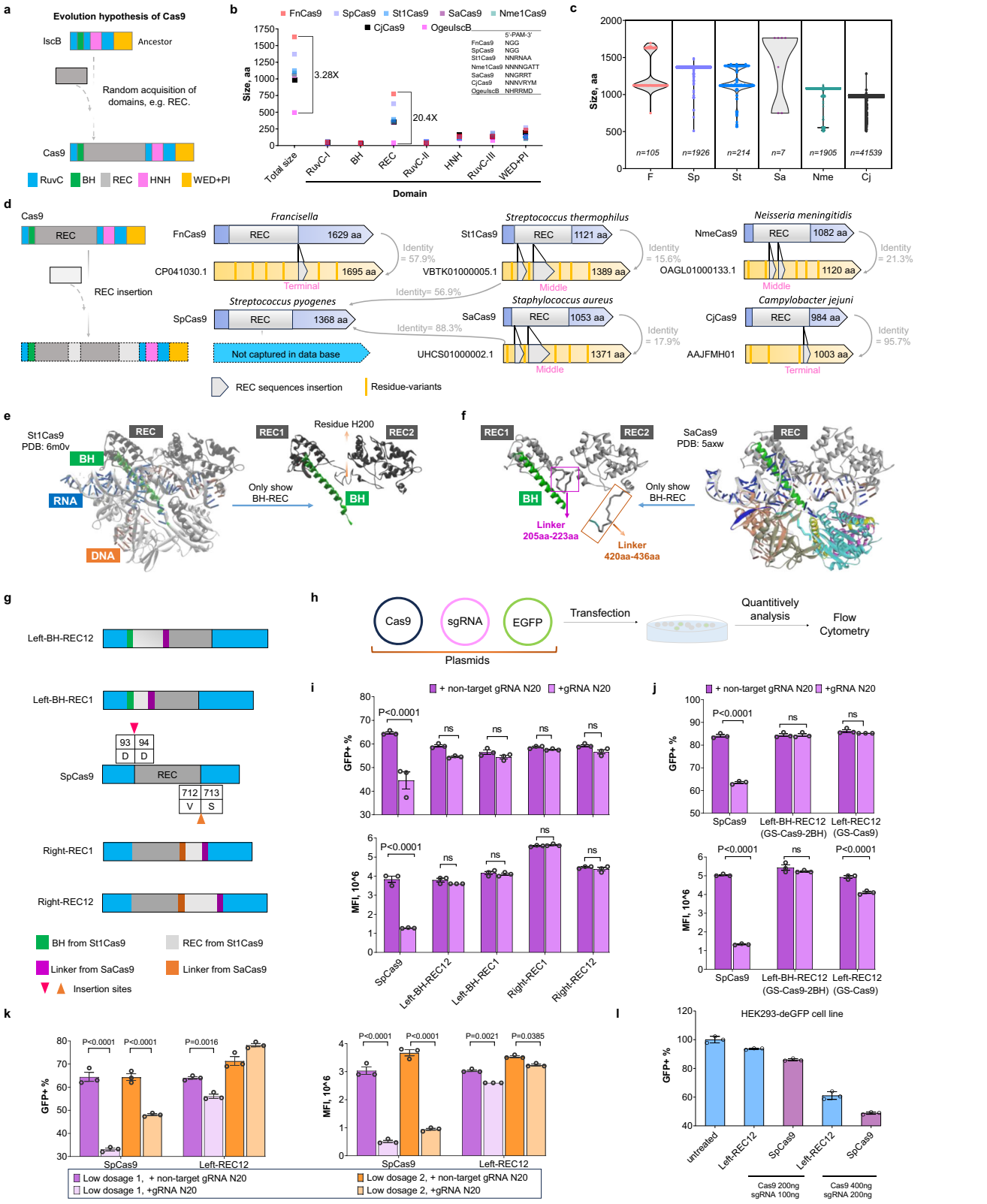
Domain expansion contributes to diversification of RNA-guided-endonucleases including Cas9. However, it remains unclear how REC domain expansion could benefit Cas9. In this study, we identify an insertion spot that is compatible with large REC insertion and succeeds in enlarging the non-catalytic REC domain of *Streptococcus pyogenes* Cas9. The natural-evolution-like giant SpCas9 (GS-Cas9) is created and shows substantially improved editing precision. We further discover that enlarging the REC domain could enable regulation of the N-terminal adenine deaminase TadA8e tethered to the Cas9 scaffold, which contributes to substantially reducing unexpected editing and improving the precision of the adenine base editor ABE8e. We provide proof of concept for evolution-inspired expansion of Cas9 and offer an alternative solution for optimizing gene editors. Our study also indicates a vast potential for engineering the topological malleability of RNA-guided endonucleases and base editors.

As a subset of RNA-guided endonucleases, the diversity of CRISPR-Cas9 orthologs provides rich material for studying its evolution[1,2]. However, CRISPR-Cas9 has not yet evolved to be an ideal scaffold for gaining additional catalytic domains[3]. The evolution hypothesis suggests that Cas9 originated from a predicted ancestor and underwent a complex evolution from small to large sizes in hundreds of millions of years, involving the expansion of various domains through random insertions[4] (Fig. 1a). Among these domains, acquisition of the recognition (REC) domain is predicted to play a critical role in enhancing the specificity of Cas9[1,4,5]. Although the topological malleability of Cas9 was realized by random insertion of non-Cas9 domain using Mu transposon[3,6], the biological influence of natural domain expansion on Cas9 remains unclear. Despite many efforts to minimize Cas9 off-target cleavage through introducing amino acid substitutions over the past decade[7–12], there is currently no direct experimental evidence supporting the hypothesis that domain expansion contributes to improving Cas9 performance. Meanwhile, compact editors have been paid excessive attention for overcoming the size limitations associated with viral-based delivery in gene editing[13,14] (Supplementary Fig. 1a). Encouragingly, lipid nanoparticle (LNP) delivery has emerged as a safer and more efficient solution for the delivery of large cargos[13,15–18], which drives us to explore various routes for engineering gene editors. Despite increasing payload size, exploring domain expansion can both aid in refining the knowledge of Cas9 evolution to give a deeper understanding of Cas9 structure, and open a new pathway to artificially increase diversity of Cas9 and develop innovative solutions for improving accuracy and precision of Cas9-based gene editors.

Here, we try obtaining knowledge from the natural evolution of Cas9 to improve the performance of gene editors. By combining bioinformatics analysis and protein engineering methods, we are inspired by nature and create a natural-evolution-like nuclease, giant SpCas9 (GS-Cas9), carrying the largest REC domain experimentally identified to date. An unreported role of domain expansion is uncovered in this study, where enlarging the REC domain could regulate the catalytic activities of the domain tethered to the N-terminus of the Cas9 scaffold. ABE8e adenine base editor shows high compatibility with the enlarged Cas9 scaffold, which results in improving the precision of base editing. As an alternative strategy, the GS-Cas9 scaffold may have huge potential to be expanded to most SpCas9-based base editors to fine-tune their performance. To

Department of Biomedical Engineering, Tufts University, Medford, MA, USA. e-mail: Qiaobing.Xu@tufts.edu

https://doi.org/10.1038/s41467-025-57154-5



our knowledge, we show a case of developing improved Cas9-based gene editors by harnessing nature-evolution-like concept that is different from the strategies reported before.

Results

REC domains of Cas9 show high flexibility in size

Zhang and colleagues predicted a total of 161,859 IsrB_IscB_Cas9 sequences from a prokaryotic database constructed by combining various databases[1], among which there are 138,334 Cas9 sequences (Supplementary Fig. 1b). However, only a minuscule fraction of these sequences carrying less than 1700aa has been experimentally characterized or structurally resolved in the past decade[19–26]. To investigate which domain exhibits the representative expansion in natural evolution, we first compared characteristics of seven proteins (one IscB and six Cas9s) structurally identified to date. Compared to an ancestor OgeuIscB, FnCas9 (size 1629aa) shows a

Fig. 1 | Comparative analysis of various Cas9 sequences and investigating REC expansion of SpCas9. a Schematic illustration of REC expansion in the Cas9-evolution-hypothesis. **b** Domain size of IscB and Cas9s. Crystal structures of all the proteins have been identified. REC, recognition (REC) lobe; BH, bridge helix; PAM, protospacer-adjacent motifs; PI, PAM-Interacting Domain. FnCas9 (PDB:5B2O), SpCas9 (PDB:4OO8), St1Cas9 (PDB:6M0V), SaCas9 (PDB:5AXW), NmeCas9 (PDB:6JDQ), CjCas9 (PDB: 6JOO), OgeuIscB (PDB:7UTN). **c** Violin plot illustrating distribution of Cas9-sizes in various species. One pot means one sequence. *n*, number of sequences in each group. **d** Highlighting evolution of Cas9 proteins in different species. F, *Francisella*; Sp, *Streptococcus pyogenes*; St1, *Streptococcus thermophilus*; Sa, *Staphylococcus aureus;* Nme, *Neisseria meningitidis* and Cj, *Campylobacter jejuni*. **e** A crystal structure of St1Cas9 (PDB:6M0V). **f** A crystal structure of SaCas9 (PDB:5AXW). **g** Schematic of REC expansion from SpCas9. Insert positions are shown in cells. **h** Workflow for testing Cas9 variants activity in HEK293T cells. Episomal EGFP plasmid was co-transfected with Cas9 and gRNA plasmids to monitor Cas9 activity. **i** Activities induced by SpCas9 and variants. Data are presented as mean ±s.d. ($n = 3$). *P* values were determined by two-way ANOVA Sidak's multiple comparisons test. **j** Testing influences of BHs with different lengths on activities. Data are presented as mean ±s.d.($n = 3$). *P* values were determined by two-way ANOVA Sidak's multiple comparisons test. **k** Variant Left-REC12 mediated EGFP disruption at different plasmid dosages. Low dosage 1, 2 ng of reporter plasmid; Low dosage 2, 4 ng of reporter plasmid. Data are presented as mean ±s.d. ($n = 3$). *P* values were determined by two-way ANOVA Sidak's multiple comparisons test. **l** Disruption activities of SpCas9 and GS-Cas9 on endogenous GFP site in HEK293-deGFP cells. Data are presented as mean ±s.d.($n = 3$). Source data are provided as a Source Data file.

3.28-fold increase in overall size and a 20.4-fold increase in REC domain size. Meanwhile, other domains increased less than 3.08-fold (Fig. 1b and Supplementary Fig. 1c). To further evaluate the distribution of REC domain sizes in large Cas9s, we filtered sequences larger than FnCas9 to get a total of 45 unique sequences (Supplementary Notes and Supplementary Table. 1) after the removal of 48 repeat sequences. These proteins were an average size of 1712aa, where CP041030.1 (size 1695aa) isolated from the *Francisella* sp. LA112445 strain carried a predicted 841aa REC indicating 22.1-fold larger than that of IscB. This REC surpasses that of CjCas9, SpCas9 and FnCas9 by 491aa, 217aa and 65aa, respectively (Supplementary Fig. 2). The analysis indicates that REC shows higher flexibility in size and tolerance for continuous insertions than other domains. On the other hand, the protospacer adjacent motif (PAM) for large Cas9 likely becomes shorter than that of compact Cas9 (Fig. 1b and Supplementary Fig. 1d). Additionally, Cas9 proteins isolated from the same genus or species show a wide range of protein size (Fig. 1c). We propose that the REC lobe may also serve as a good mediator for matching and balancing protein-size expansion during the evolutionary process of Cas9s, thereby maintaining the functional conformation of multi-domain[9,27,28].

To further explore possible insertion sites for REC expansion, we then systemically analyzed 45,696 Cas9 sequences isolated from five kinds of organisms (genus *Francisella*, species *Streptococcus pyogenes*, *Streptococcus thermophilus, Staphylococcus aureus, Neisseria meningitidis* and *Campylobacter jejuni*, in which at least one Cas9 has been structurally identified) by removing repeats. Except for *S. pyogenes* Cas9 sequences, we found inspirations of REC lobe expansion. Likewise, the predicted protein UHCS01000002.1 isolated from *S. aureus* is 318aa longer than SaCas9 and contains a predicted REC that is 268aa longer than that of SaCas9. Insertions mainly appear in either the middle (for *S. thermophilus*, *S. aureus*, *N. meningitidis* and *C. jejuni*) or boundaries (for *Francisella, S. thermophilus* and *S. aureus*) of REC domain (Fig. 1d, and Supplementary Figs. 3–8). For the large Cas9 in *Francisella*, the expansion seems easier to occur at the C-terminus of REC. These positions might be used as sites for REC insertion. Most notably, UHCS01000002.1 shows an 88.3% pairwise sequence identity with SpCas9 (Supplementary Fig. 9), while LR822033.1 and VBTK01000005.1 isolated from *S. thermophilus* have 99.4% and 56.9% sequence identity with SpCas9, respectively (Supplementary Fig. 10). Since FnCas9 and SpCas9 were younger than StCas9, SaCas9, and NmeCas9[29], we suppose that SpCas9 might have evolved from SaCas9 or St1Cas9 through domain expansion. Though we didn't capture expansion sequences derived from SpCas9, our analysis indicates that SpCas9 may have high compatibility with the REC domain of SaCas9 or St1Cas9.

Building giant SpCas9 by enlarging REC domain

Based on the bioinformatic analysis above, we assumed that REC domains have the ability to tolerate large-size insertions. SpCas9 has been extensively studied in terms of its crystal structure and mechanism of function, which enables it to be the most widely used in both academic and industrial fields[13]. Based on these inspirations above, we tried to explore whether we could create giant SpCas9 variants by enlarging the REC domain and gaining biological insights[30]. Multiple-sequence alignments of *S. thermophilus* and *S. aureus* did not show conserved motifs near the boundaries of REC domains (Supplementary Figs. 5, 6, 11), which suggests that these sites might serve as recombination hot spots compatible with foreign insertions.

Because domain combinations are often found in only one sequential order in the evolution of the majority of multi-domain proteins[31,32], we first tried inserting the REC lobe of St1Cas9 at the termini of the REC domain in SpCas9 to generate expansion variants driven by these understandings. Given flexible linker regions in Cas9 appear to play a role in the inactive-to-active conformational transition of multiple domains[19,20,33], we selected two short-flexible-linkers from SaCas9 (residues T205 to D223 or K420 to T436) to link insertions without using any non-Cas9 sequences (Fig. 1e and f). We designed and constructed four variants in which different REC elements were optimized for humans using codon usages from IDT (Fig. 1g). To avoid Cas9 variants failing to function in mammalian cells, we chose human embryonic kidney (HEK) 293T cells to evaluate the activity of our variants with an EGFP plasmid interference assay modified from previous research[34] (Fig. 1h). Unfortunately, we did not observe EGFP disruption activity from any of the four variants (Fig. 1i). We speculated that enlarged Cas9-variants might also not need extended BH domains. We also observed that the excessive length of the BH domain reduced the cleavage activity of SpCas9-2BH (Supplementary Fig. 12a). Therefore, we removed the St1-BH sequence in the variant Left-BH-REC12 to generate the Left-REC12 variant. The new variant still did not induce reduction of GFP^+ cell populations, but did lead to a 20% decrease in mean fluorescence intensity (MFI) of GFP^+ cells (Fig. 1j). Excitingly, we observed a significant decrease in GFP^+ cells in presence of the Left-REC12 variant when the substrate plasmid concentration was reduced to a lower level (Fig. 1k). Then, we tried to use a human cell line bearing an integrated construct that constitutively expresses a deGFP protein, HEK293-deGFP[35], to evaluate Left-REC12 activity on chromosome sites. We observed 45.5% to 70% GFP disruption activity relative to SpCas9 in the presence of various doses of the Left-REC12 variant (Fig. 1l). We speculated that domain expansion might impact Cas9 kinetics which caused the great differences of EGFP disruption in presence of different molar concentrations of DNA substrate[36,37]. Then, we specially named Left-REC12 as giant SpCas9 (GS-Cas9). The GS-Cas9 protein has a length of 1780 aa and a 207 kDa molecular weight, which is 1.3 times larger than SpCas9 and might form an unbalanced bilobed structure (Fig. 2a, Supplementary Fig. 12b). The REC domain of GS-Cas9 consists of 1036 amino acids making it 1.66-fold larger than that of SpCas9. This represents not only the largest Cas9 protein size but also the largest REC domain size experimentally characterized to date. Though others have built synthetic Cas9 scaffolds using an engineered Mu transposon system[6,38], all these efforts introduced Mu-recognized sequences or non-Cas9 domains into variants. By contrast, GS-Cas9 is a nature-like Cas9 without any non-Cas9 domains being introduced. We also

https://doi.org/10.1038/s41467-025-57154-5

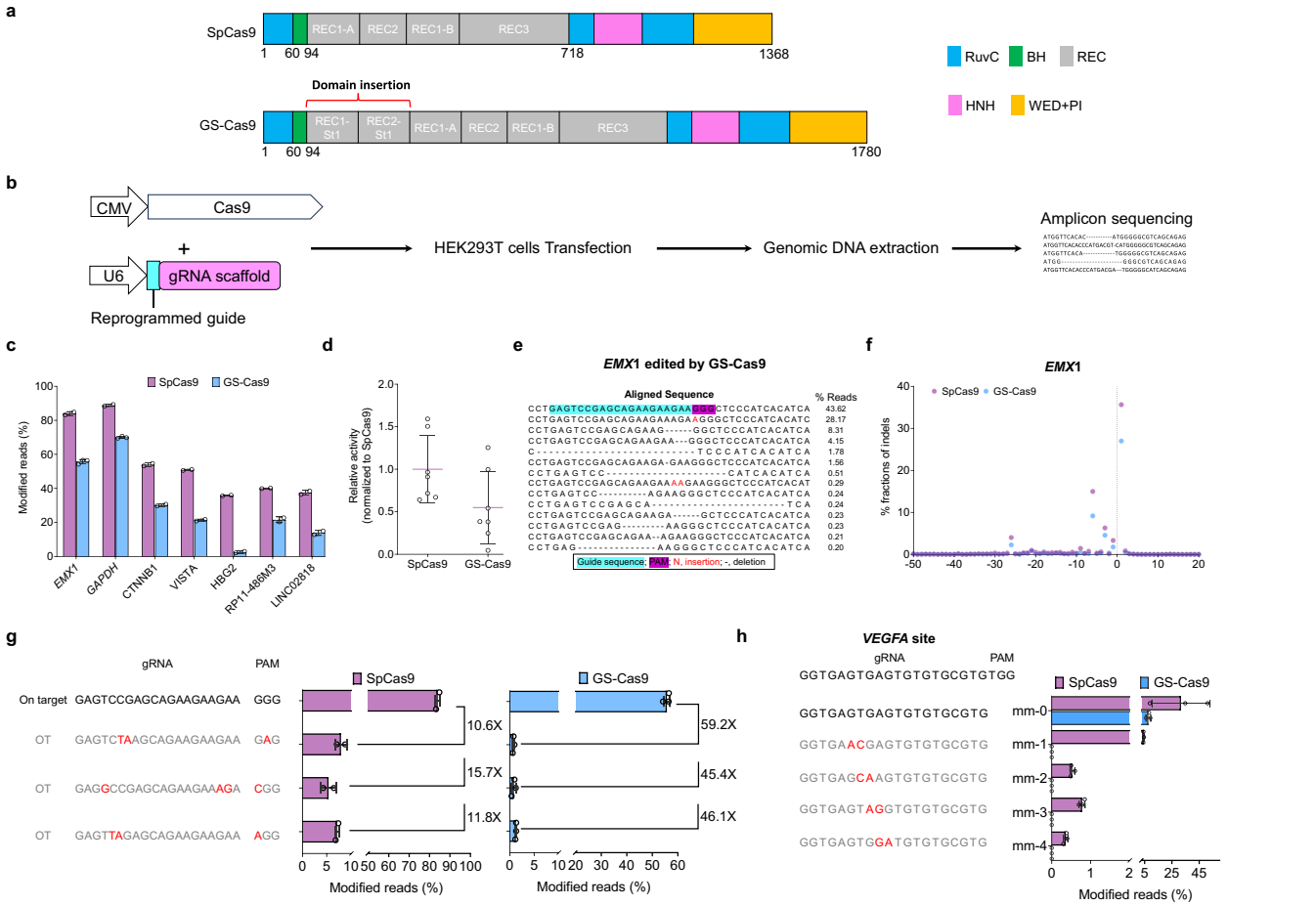


Fig. 2 | Investigating GS-Cas9 performance on human genome. a Domain organization of SpCas9 and GS-Cas9. **b** Workflow for systemically testing GS-Cas9 activity in HEK293T cells. **c** Edits rates generated by SpCas9 and GS-Cas9 at seven loci in HEK293T cells. Data are presented as mean ±s.d. (SpCas9, $n = 2$; GS-Cas9, $n = 3$). **d** Normalized editing frequencies for 7 target sites for SpCas9 and GS-Cas9. Each dot represents a different guide. Each data point represents the relative activity at each site calculated from (**c**). **e** Representative sequences of the human *EMX1* site targeted by GS-Cas9. sgRNA target site and PAM are indicated by blue and purple respectively. **f** Average indel length at *EMX1* site from (**c**). **g** Targeted deep-sequencing analysis of off-target sites for the *EMX1* site. Data are presented as mean ±s.d. (SpCas9, $n = 2$; GS-Cas9, $n = 3$). **h** Modified levels with guide sequences containing double-base mismatches at *VEGFA* site. Data are presented as mean ±s.d. ($n = 3$). Source data are provided as a Source Data file.

demonstrate an insertion site unidentified by transposon[6,38], which may suggest different compatibilities of SpCas9 with non- or nature-Cas9 domains. Importantly, SpCas9 and GS-Cas9 can allow for comparative study to deepen the understanding of influences induced by domain expansion on Cas9 properties.

We next evaluated the expression of GS-Cas9 using western blotting. The lower disruption efficiency of GS-Cas9 may be attributed in part to lower expression relative to that of SpCas9. However, SpCas9-2BH, GS-Cas9 showed similar expression levels (Supplementary Fig. 12c). Deletions or mutations of REC domain in various Cas9 nucleases generally led to reduced protein expression in human cells[19,26], which suggests that topological changes may play a critical role in Cas9 expression or stability.

Enlarging REC lobe could enhance editing precision on human genome

To test whether the enlarged RNA-guided nuclease could enhance editing precision, we investigated the editing performance of GS-Cas9 on seven endogenous genome loci in HEK293T cells (Fig. 2b). For wild type SpCas9, modified levels ranged from 35.8%-88.8% at the seven genomic loci tested, whereas for GS-Cas9 modified levels ranged from 2.6%-70.3% (Fig. 2c). Totally, GS-Cas9 displayed 55% on-target activity of SpCas9 across seven sites tested (Fig. 2d). Alleles showed patterns of substitutions, insertions, and deletions (Fig. 2e, Supplementary Fig. 13), however, we did not observe different indel patterns between SpCas9 and GS-Cas9 (Fig. 2f). Next, we tested the top 3 known genomic off-target (OT) sites for *EMX1* editing as identified by GUIDE−seq (genome-wide, unbiased identification of double-strand-breaks enabled by sequencing)[7]. At all three off-target-sites, we observed lower editing levels mediated by GS-Cas9 (averaging 0.94%, 0.83% and 1.24% at these 3 off-target sites, respectively) than that of SpCas9 (averaging 7.95%, 5.37% and 7.12%, respectively). Additionally, the ratio of on-target to off-target editing increased from an average 12.7 for SpCas9 to 50.2 for GS-Cas9 across these 3 off-target sites (Fig. 2g). To further evaluate the tolerance of GS-Cas9 for mismatched target sites, we chose four mutated guide sequences for the *VEGFA* site introducing double-base mismatches at different PAM-distal positions reported in previous research[8]. Compared with SpCas9, GS-Cas9 induced undetectable modified reads with all mismatched guides, while SpCas9 induced 0.36%-4.21% of modified reads (Fig. 2h). Two sites contain either a cytosine-rich homopolymeric sequence or a sequence with multiple TG repeats, where GS-Cas9 mediated significantly lower editing compared to SpCas9 (14-fold lower at the *HBG2* site and 4-fold lower at the *VEGFA* site, respectively).

This data demonstrated that GS-Cas9 improved targeting specificity and retained detectable activity. FnCas9 possesses higher

specificity than SpCas9[39], which may be attributed in part to its larger REC lobe. Excessive expression of gene editors could lead higher off-target edits[40], domain expansion may improve editing accuracy by partially altering protein expression of Cas9. Engineered high-fidelity SpCas9 variants always suffer loss of editing activity compared to wild type, for instance, HypaSpCas9 demonstrates 60% activity relative to SpCas9[12,41,42]. GS-Cas9 shows 55% relative activity compared to SpCas9 across 7 tested sites which is in the similar activity level to HypaSpCas9 in HEK293T cells.

Enlarging REC domain reduces RNA editing of adenine deaminase TadA8e in base editor

In contrast to SpCas9 and SaCas9, the RuvC domain interacts with the REC domain in FnCas9, a naturally occurring large Cas9[21] (Fig. 3a). We also observed that REC-domain sizes often increase with increases in other domain sizes of Cas9s (Fig. 1c and Supplementary Fig. 1). For instance, the SpCas9-RuvC domain (307 residues) is 1.45-fold larger than that of SaCas9 (212 residues). We hypothesized that REC expansion may induce extra interaction between domains than the parent. Then we chose the ABE8e editor as a fusion model of Cas9 for gain-of-function to investigate whether REC expansion could impact N-terminal catalytic domain activities. ABE8e fusion contains an adenine deaminase mutant TadA8e (166aa) with higher activity fused to the N-terminus of the nickase SpCas9 (D10A), in which TadA8e is exposed to the environment with resulting in high mobility and no specific interaction with the SpCas9 scaffold[43,44] (Fig. 3a). The exposure increases the freedom of TadA8e, thereby enhancing its chances of interacting with other DNA or RNA nucleotides, resulting in a higher occurrence of Cas9-independent off-target editing events[45–47]. Cas9-independent off-target DNA editing and RNA editing activities of ABE are positively correlated with the activity of the tethered adenine deaminase[44,48,49]. We proposed a hypothetical model in which REC expansion may lead to a rearrangement of the original REC domain and shorten the distance between REC and N-terminal TadA by enlarging the coverage of the engineered REC lobe, thereby generating influences on TadA (Fig. 3b).

We first established a time-saving and cost-effective method, plasmid-based RNA editing and DNA editing reporter assay (pbREADER), to evaluate editing activities of ABE8e variants (Fig. 4a, Supplementary Figs. 14 to 15, Supplement Notes). Four variants ABE8e-2BH, GS-ABE8e, GS-ABE8e-1.5BH and GS-ABE8e-2BH were constructed (Fig. 4b). We observed that GS-ABE8e-2BH induced comparable RNA editing and DNA editing activities with ABE8e, but MFI reduction of DNA editing mediated by GS-ABE8e-2BH was 2.3-fold less compared to ABE8e (Supplementary Fig. 16a and b). Meanwhile, we did not observe both RNA editing and DNA editing activities from the ABE8e/RR12 variant with the REC domain inserted at the C-terminus of Cas9 scaffold in ABE8e. Given the SpCas9/Right-REC12 data, we suggest that the insertion at the C-terminus of the REC domain may lead to protein misfolding. GS-ABE8e-2BH showed efficient editing activities, which indicates that GS-Cas9-2BH retains on-target binding. As a long α-helix, BH also acts as a rigid spacer between REC and RuvC-I of Cas9[50]. Compared to GS-ABE8e-2BH, truncated BH induced not only a decrease in RNA editing but also increased DNA editing activity (Fig. 4c-f). Of note, EGFP activation of GS-ABE8e mediated by RNA editing reduced 4.46-fold and 3.64-fold compared to the GS-ABE8e-2BH and ABE8e, respectively. We observed similar trends in transfection with or without loading non-target sgRNA containing 20nt-guides. We did not observe differences in MFI mediated by RNA editing for these variants. However, MFI difference induced by DNA editing from GS-ABE8e showed an almost 2-fold increase relative to GS-ABE8e-2BH (Fig. 4g). This indicates that GS-ABE8e maintains the same DNA editing activity on EGFP plasmid as ABE8e. We also observed similar differences between ABE8e and ABE8e-2BH (Supplementary Fig. 16c). Like GS-Cas9, ABE8e-expansion variants with the single-BH configuration exhibited the best performance. This aligns with the natural evolutionary selection of Cas9s, in which near-fixed BH length enables adaptation to wide range of Cas9 sizes. Unlike GS-Cas9, western blotting indicated that GS-ABE8e, GS-ABE8e-1.5BH and GS-ABE8e-2BH showed similar expression levels with ABE8e (Supplementary Fig. 16d). We further assessed whether GS-ABE8e reduced off-target RNA editing activity on endogenous transcripts within the cell. After transfection of

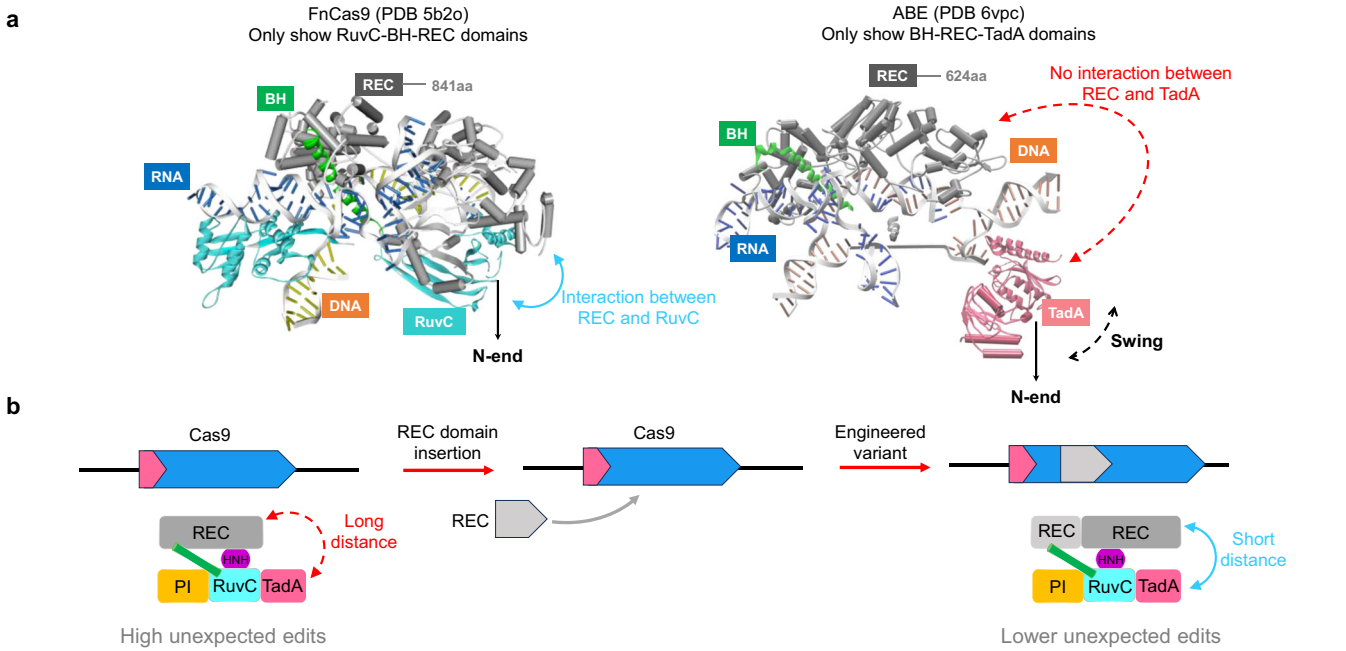


Fig. 3 | ABE8e was chosen for investigating the influences of REC expansion on N-end catalytic domain. a Protein structure of FnCas9 and ABE8e. Left, FnCas9 (PDB 5b2o); right, ABE8e (PDB 6vpc) contains an engineered deoxyadenosine deaminase TadA from *Escherichia coli* and a nickase SpCas9 (D10A). **b** Hypothetical model of REC expansion for Cas9 fusion. In this proposed model, the insertion may lead to a rearrangement of the original REC domain and shorten the distance between REC and TadA by enlarging the coverage of the engineered REC lobe, thereby generating additional interactions between these two domains.

https://doi.org/10.1038/s41467-025-57154-5

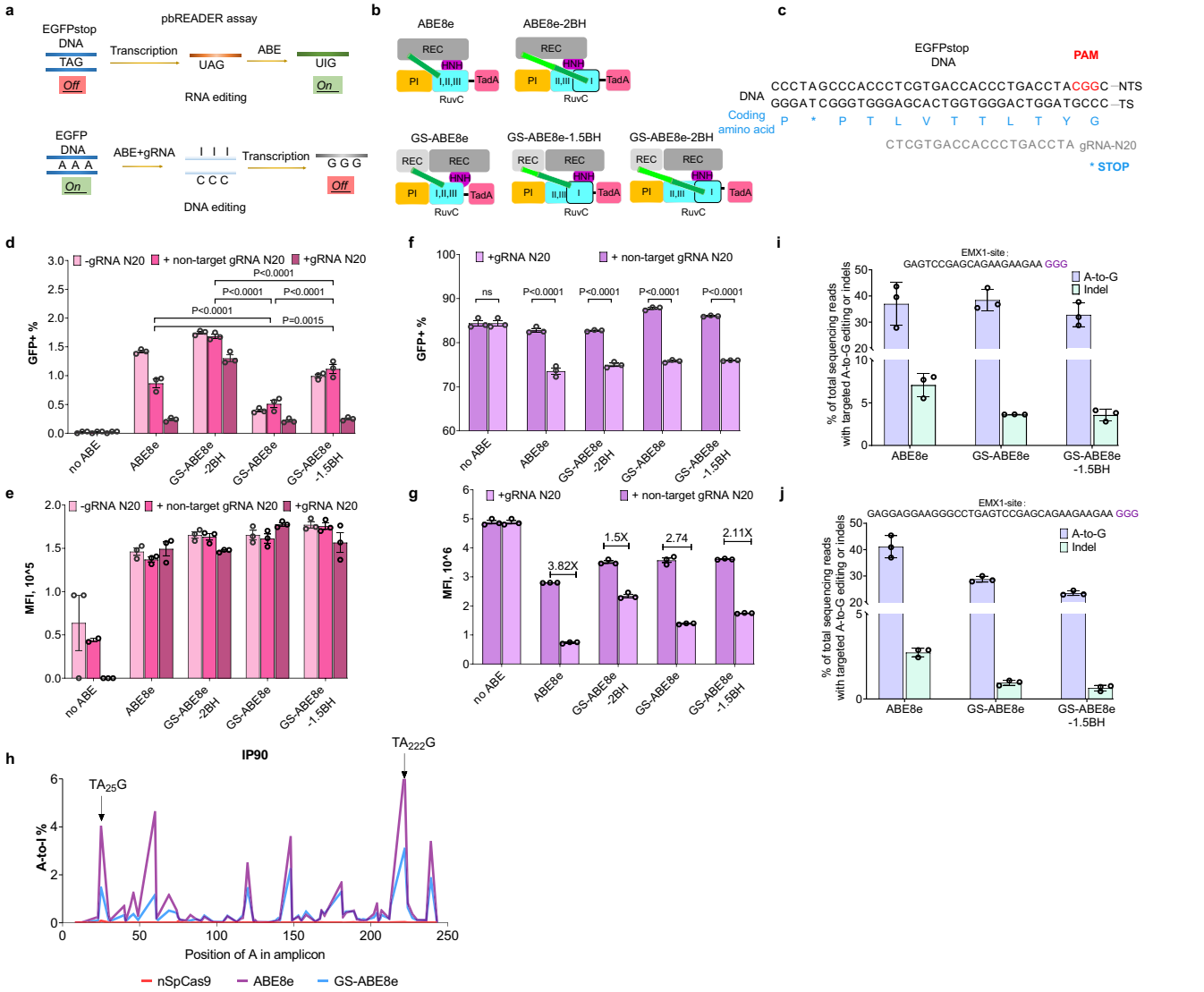


Fig. 4 | REC expansion could regulate the catalytic domain fused to N-end of SpCas9. a An illustrating work-process illustration of the plasmid-based RNA editing and DNA editing reporter (pbREADER) assay. **b** A schematic illustration of the ABE8e and hypothetical GS-ABE8e variants. **c** Representative sequences of reporter plasmid pCMV-EGFP(W58stop) targeted by SpCas9 or GS-Cas9. Below, guide sequences with 20-nt length. NTS, non-target strand; TS, targeted strand. **d, e** Activation of EGFP mediated by various base editors. Data are presented as mean ±s.d. ($n = 3$). **f, g** Disrupting EGFP with DNA editing mediated by various base editors. Data are presented as mean ±s.d. ($n = 3$). *P*-values were determined by two-way ANOVA Tukey's multiple comparisons test. **h** Average A-to-I RNA editing frequencies by nSpCas9, ABE8e and GS-ABE8e mutants among 77 adenosines in IP90 mRNA transcripts. Data are presented as mean ±s.d. ($n = 3$). **i, j** Cumulative A•T to G•C base editing efficiencies and indel formation at the human genomic *EMX1* site in HEK293T cells treated with various base editors loaded with 20nt or with 35nt guide RNA. Data are presented as mean ±s.d. ($n = 3$). Source data are provided as a Source Data file.

HEK293T cells by nickase SpCas9 (D10A), ABE8e, and GS-ABE8e, RNA was extracted from cells. The IP90 transcript reported in previous study[49,51] was used to evaluate off-target RNA editing. The amplicon was produced by PCR after complementary DNA (cDNA) and analyzed for A-to-I editing by high-throughput sequencing. We found that GS-ABE8e drastically reduced A-to-I levels at all A positions inside the amplicon (Fig. 4h). Especially, A-to-I efficiencies at TAG motif reduced more than 2-fold compared to ABE8e. These results reveal that enlarging the REC domain succeeds in regulating adenine deaminase activities in ABE8e.

We then evaluated whether GS-ABE8e variants still could retain on-target editing activity on a genome locus in HEK293T cells. We observed similar A-to-G editing rates at the *EMX1* site for these variants loaded with 20nt-gRNA (cumulative editing rates of 36.98%, 38.44% and 32.79% for ABE8e, GS-ABE8e, and GS-ABE8e-1.5BH, respectively) (Fig. 4i and Supplementary Fig. 17). Importantly, we observed decreases in indel levels for both REC expansion variants (averaging indels of 7.09%, 3.63%, and 3.58% for ABE8e, GS-ABE8e, and GS-ABE8e-1.5BH, respectively). Enlarged REC domains may provide protection for the non-target strand by reducing the ssDNA exposed to solvent, which may decrease double-strand breaks and indels. REC expansion variants induced slightly lower A-to-G efficiencies in the presence of a 35nt-gRNA (cumulative editing rates of 41.10%, 28.78% and 23.50% for ABE8e, GS-ABE8e, and GS-ABE8e-1.5BH, respectively), while indel levels were around 4.25-fold lower compared to ABE8e (2.72%, 0.95%, and 0.64% for ABE8e, GS-ABE8e, and GS-ABE8e-1.5BH, respectively) (Fig. 4j and Supplementary Fig. 18). We did not observe edits outside of the 20-nt protospacer, which suggests that enlarging the REC lobe does not enlarge the editing window in presence of longer guide sequences.

Here we show that enlarging the REC domain enables Cas9 scaffold to reduce the activities of N-end fused catalytic domain

in ABE8e on RNA transcripts, which indicates an unreported role of REC domain expansion. Given the non-specific DNA mutations mediated by free HNH or RuvC nucleases[52,53], we suggest that REC acquisition or expansion might play a 'peacemaker' role during natural evolution to influence catalytic domains to work within a specific space and reduce non-specific cleavage (Fig. 3b). GS-Cas9 variant exhibited advantages in editing precision over SpCas9, therefore we performed all subsequent experiments with GS-ABE8e to systemically compare it with ABE8e and determine the characteristics of GS-ABE8e.

Enlarging REC domain contributes to reducing ABE8e unexpected editing in plasmid-based reporter system

The orthogonal R-loop assay is commonly used for evaluating Cas9-independent off-target editing induced by base editor[54]. Because the GS-Cas9 scaffold was a chimera of multiple Cas9s, we then investigated whether GS-ABE8e was compatible with gRNA-scaffolds from related Cas9 systems using the EGFP disruption assay (Supplementary Fig. 19a and b). GS-ABE8e still exhibited the highest GFP disruption activity in the presence of a SpCas9 gRNA-scaffold, which was like ABE8e. We observed similar activities for ABE8e and GS-ABE8e in the presence of gRNAs with Sa- or St1-gRNA-scaffolds. Interestingly, GS-ABE8e showed greater activity than ABE8e in the presence of a gRNA-FnCas9-scaffold (Supplementary Fig. 19c). Additionally, GS-ABE8e did not show activity with a gRNA carrying the SaCas9-gRNA-scaffold and a guide sequence targeting an EGFP-DNA region bearing the promiscuous 5′-CGGAGT-PAM for SaCas9 and SpCas9 (Supplementary Fig. 19d and e). These results indicate that GS-Cas9 scaffold retains its original orthogonality.

sgRNA may suffer degradation from nuclease in cells[55], we next evaluated the tolerance of GS-ABE8e for truncated guide sequences. Compared with ABE8e, GS-ABE8e exhibited lower relative activity with 12nt, 10nt and 8nt guides (Fig. 5a), which indicates that GS-ABE8e is more sensitive to truncated guide sequences. To further determine whether GS-ABE8e is still sensitive to mismatched target sites, we systematically mutated the EGFP guide sequence to introduce double-base mismatches at various positions (Fig. 5b). Compared with ABE8e, GS-ABE8e showed lower relative activity with mismatches located inside of the 1- to 10-base pair seed sequence. This revealed that enlarging REC domain also enhanced targeting specificity of ABE8e base editor.

Base editors could generate unexpected ultra-low edits outside of the protospacer sequence (that is, out-of-protospacer), which is lower than the detection limit of next generation sequencing (NGS)[56]. The EGFP reporter enables a much lower 10-fold detection limit for base editing than NGS-based strategies[57]. To evaluate whether GS-ABE8e has the potential to reduce edits out-of-protospacer on non-target strands, we compared the EGFP reporter activities of GS-ABE8e and ABE8e on an adenine base located at position 4 out-of-protospacer (that is, A (-4)). Notably, we observed that A (-4) out-of-protospacer edits for GS-ABE8e reduced 13.3-fold compared to ABE8e (Fig. 5c and Supplementary Fig. 20). In an orthogonal R-loop assay using reporter plasmid, GS-ABE8e mediated EGFP activation was 6.5-fold lower than ABE8e (Fig. 5d), which contributes to higher resolution outcomes on low EGFP-DNA-substrate concentrations (Supplementary Fig. 21). Moreover, in an orthogonal R-loop assay for Cas9-independent off-target DNA editing on genome (Fig. 5e), GS-ABE8e expectedly generated much lower rates of off-target effects (averaging 0.29% and up to 0.56%), but ABE8e generated much higher Cas9-independent off-target editing (averaging 3.1% and up to 6.8%) (Fig. 5f). Reasonably, this improvement might benefit from extra interaction between the enlarged REC domain and TadA domain. Meanwhile, GS-ABE8e retains similar editing activity with ABE8e on locus site 1 (Fig. 5g).

Anti-Cas9 protein AcrIIA4 can inhibit ABE activity by occupying both PAM-interacting and non-target DNA strand cleavage catalytic pockets of SpCas9[58]. In our EGFP disruption assay, we observed no MFI reduction induced by GS-ABE8e in the presence of AcrIIA4 (MFI reduction 24% for ABE8e vs 0% for GS-ABE8e, respectively) (Supplementary Fig. 22). These data reveal that GS-ABE8e shows greater sensitivity to anti-Cas9 protein.

GS-ABE8e shows lower indels and Cas9-dependent off-target editing than ABE8e on human genome

It has been challenging to achieve a balance between the target and off-target activities of ABE in previous studies. Though Cas9-independent RNA off-target editing and DNA off-target editing of ABE8e has been reduced through directed evolution of TadA8e, TadA8e mutants were difficult to reduce unexpected indel events[44,48]. Conversely, ABE8e constructs were compatible with previously described high-fidelity SpCas9 variants (e.g., SpCas9-HF1, evoSpCas9, SuperFi-Cas9) bearing different residue-mutations and succeeded in minimizing Cas-dependent off-target editing, which unfortunately showed substantial loss in on-target editing[59–62]. To investigate the performance of GS-ABE8e more thoroughly, 11 endogenous human genome loci previously reported[44,48,63,64] were tested. We compared ABE8e with GS-ABE8e side by side using 11 additional gRNA sites. The activity (defined as the editing level at the position with the highest A-to-G rate in each site) of GS-ABE8e ranged from 40.4% to 81.0%, which was similar with ABE8e (45.6% to 74.7%) at 10 of 11 sites (Fig. 6a and b). Compared with ABE8e, indel levels induced by GS-ABE8e at 11 target sites decreased by 1.7- to 3.4-fold (0.46% to 2.49% for GS-ABE8e vs 1.07% to 8.63% for ABE8e, respectively) (Fig. 6c). Across all 12 tested sites (*EMX1* site in Fig. 4i and 11 sites in Fig. 6c), GS-ABE8e induced a 2-fold decrease of indels compared to ABE8e (Fig. 6d). We examined the base editing window of GS-ABE8e and ABE8e. Consistent with ABE8e, GS-ABE8e can efficiently edit A2-A8 positions, counting the PAM as positions 21-23 (Fig. 6e). Because REC expansion enhanced targeting specificity of SpCas9, we further selected Cas9-dependent off-target sites reported in previous study[44,63] to investigate off-target activity of GS-ABE8e in HEK293T cells. The top 5 or 2 known ABE off-target loci for *EMX1* and ABE site 2 (*VISTA* enhancer) editing were evaluated. We observed a decrease in editing at one of the top two off-target sites for ABE site2 when comparing GS-ABE8e to ABE8e (0.95% vs 2.70%) (Fig. 6f). At the top 5 *EMX1* off-target sites, we observed remarkable decreases in GS-ABE8e-mediated A-to-G editing (averaging 44.63% vs 31.1%, 1.28% vs 0.60%, 4.65% vs 2.58%, 2.19% vs 1.37%, 0.1% vs 0.00% at the top five sites for ABE8e and GS-ABE8e, respectively) (Fig. 6g). The ratio of on-target to off-target editing was much higher than ABE8e at these sites (Fig. 6h and i). GS-ABE8e also produced lower rates of indels at all off-target sites (averaging 2.17% vs 1.04%, 0.13% vs 0.01%, 0.11% vs 0.06%, 0.21% vs 0.12%, 0.007% vs 0.005%, at the top 5 sites ABE8e and GS-ABE8e, respectively) (Fig. 6j and k). GS-ABE8e also mediated lower A-to-G edits at 19 out of 20 positions on non-target strands across 12 sites (only positions with editing >0.1% were counted for this conclusion) (Supplementary Fig. 23). These observations reveal that GS-ABE8e is a highly efficient adenine base editor with reducing unexpected indels and off-target edits.

TadA8e is reported with the much higher deoxyadenosine deaminase activity and leads higher off-target editing (e.g., Cas9-dependent, Cas9-independent, and RNA transcriptome) than that of variants with lower activity[44]. GS-Cas9 allows base editors to reduce unexpected edits even when carrying TadA8e. Importantly, our study demonstrates that the same goal can be achieved by adjusting the topological malleability of the Cas9 scaffold, unlike previous research which aimed to reduce unwanted edits through continuous mutagenesis of TadA8e[51]. GS-Cas9 has the potential to balance high on-target editing with low off-target editing and off-targets for therapeutic applications of base editors containing catalytic domains tethered to the N-terminus in future research. Here, our data demonstrate advantages of the GS-Cas9 scaffold in improving overall performance of adenine base editor.

https://doi.org/10.1038/s41467-025-57154-5

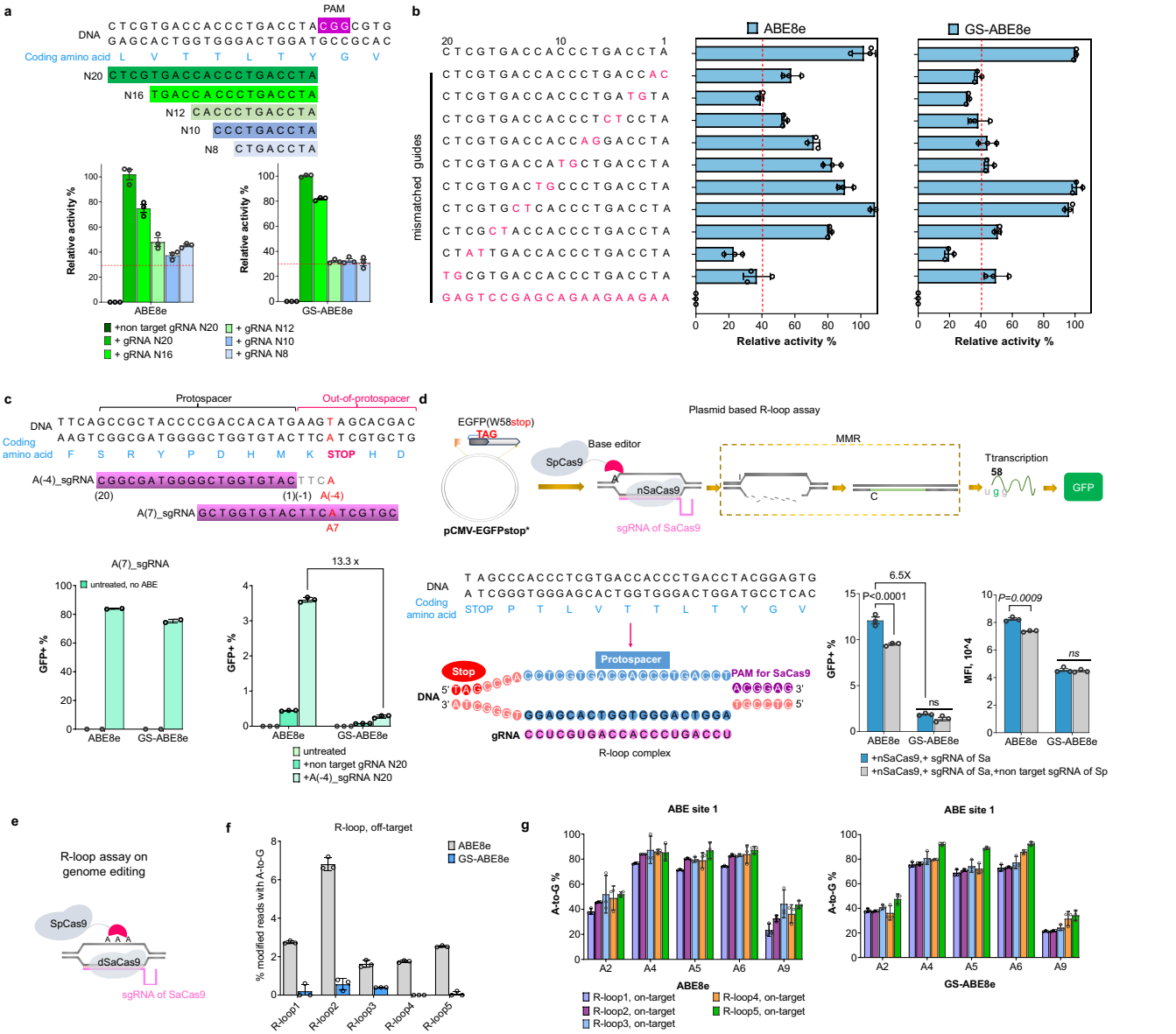


Fig. 5 | Investigating performance of GS-ABE8e using a plasmid-based reporter system. a Targeting ABE8e and GS-ABE8e to a representative EGFP locus in HEK293T cells with gRNAs containing guides in various lengths. Data are presented as mean ±s.d. ($n = 3$). **b** EGFP disruption with guide sequences containing consecutive transversion mismatches. Data are presented as mean ±s.d. ($n = 3$). **c** Investigating base editors' activities on out-of-protospacer A-to-G editing. 10 ng of EGFP(Q81stop) plasmid was used for each well. Data are presented as mean ±s.d. ($n = 2$ or 3). **d** Plasmid-based orthogonal R-loop assay. Here, EGFPstop was activated by Cas9-independent off-target DNA editing. Data are presented as mean ±s.d. ($n = 3$). *P* value was determined by two-way ANOVA Sidak's multiple comparisons test. NS, not significant. **e** Orthogonal R-loop assay overview on human genomic loci. **f** Cas9-independent off-target A•T-to-G•C editing frequencies detected by the orthogonal R-loop assay at each R-loop site with dSaCas9 and a SaCas9 sgRNA. Each R-loop was performed by co-transfection of ABE8e or GS-ABE8e, and a SpCas9 sgRNA targeting site ABE site1 with dSaCas9 and a SaCas9 sgRNA targeting R-loops 1–5, respectively. Amplicons of R-loops sites were analyzed by high-throughput sequencing. For all plots, bars represent mean ±s.d. of three independent biological replicates. **g** On-target base editing efficiencies for ABE8e and GS-ABE8e in HEK293T cells at ABE site 1 in HEK293T cells for the orthogonal R-loop assay. Amplicons of ABE site 1 were analyzed by Sanger sequencing. For all plots, dots represent individual biological replicates and bars represent mean ±s.d. of three independent biological replicates. Source data are provided as a Source Data file.

GS-Cas9 scaffold is compatible with various TadA8e variants

Different substitutions introduced in deaminase can critically affect its compatibilities with Cas homologs[44]. We then extensively investigated whether various TadA8e-derived variants could be compatible with the GS-Cas9 scaffold. We first tested the possible compatibility of GS-Cas9 with other TadA9 variants (with N108Q and L145T variants in Tad8e)[48], GS-ABE9 retained slightly lower MFI disruption activities than ABE9 (Supplementary Fig. 24). After evaluation at 9 endogenous sites, we observed that both ABE9 and GS-ABE9 showed moderately lower activity than GS-ABE8e, but GS-ABE9 significantly reduced unexpected indel levels at tested sites and narrowed the editing window from 7 nucleotides (A2 to A7) to 3 nucleotides (A4 to A6) (Supplementary Fig. 25). This data shows that specific residues can impact the compatibility of TadA with GS-Cas9 scaffold. Additionally, GS-Cas9 also showed good capabilities with TadA8e (N46L) and TadA8e (V106W) variants previously reported[51,65], which showed higher editing efficiencies than GS-ABE9 editor (Supplementary Fig. 26). Excitedly, we observed that GS-ABE8e(V106W) reduced 3-fold EGFP activation mediated by RNA editing than ABE8e(V106W) (Supplementary Fig. 27). This data further suggests that previous strategies reported for

https://doi.org/10.1038/s41467-025-57154-5

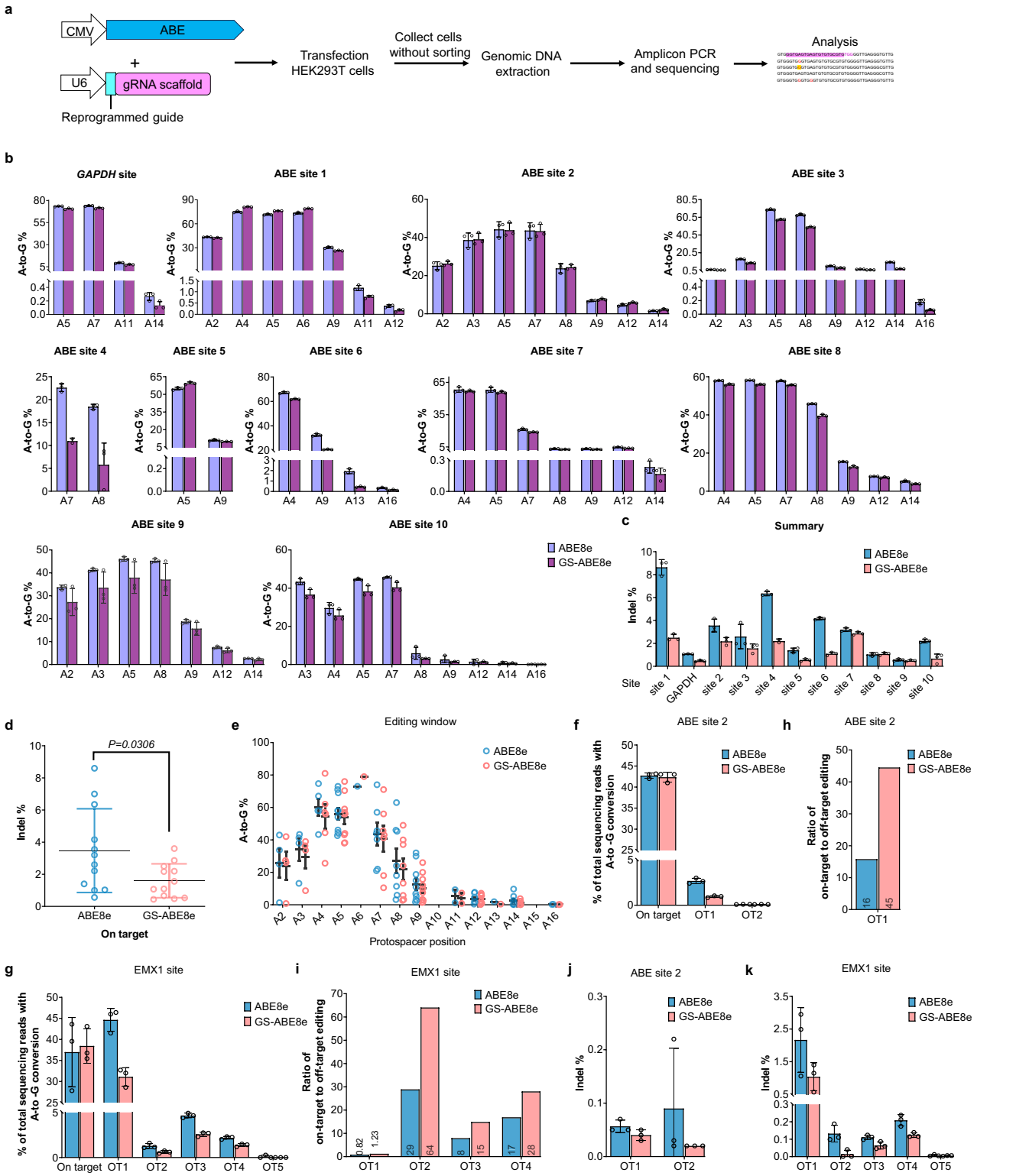


Fig. 6 | Systematically comparing ABE8e and GS-ABE8e gene editing performance on human genomic loci in HEK293T cells. a Workflow for testing GS-ABE8e in HEK293T cells. **b** Evaluation of the A-to-G editing efficiencies of ABE8e and GS-ABE8e at 11 representative endogenous genomic sites in HEK293T cells. Data are presented as mean ±s.d. ($n = 3$). **c** Indel formation in HEK293T cells treated as described in (**b**). Data are presented as mean ±s.d. ($n = 3$). **d** Comparison of indels induced by ABE8e and GS-ABE8e at 12 target sites including 11 sites in (**c**) and EMX1 site in Fig. 4i. Each data point represents the average indel frequency at each site calculated from 3 independent experiments. *P*-values were determined by two-tailed student's *t*-test. **e** Frequencies of A-to-G editing by ABE8e or GS-ABE8e editor across the protospacer positions 1-20 from (**b**). **f** Cas9-dependent DNA off-target analysis comparing ABE8e and GS-ABE8e at ABE site 2. **g** Cas9-dependent DNA off-target analysis comparing ABE8e and GS-ABE8e at *EMX1* site. **h**, **i** On-target:off-target editing ratios for two sites in (**f**) and (**g**). For all plots, bars represent mean values, and error bars represent the s.d. of three independent biological replicates. **j** Indel formation at two off-target sites in (**f**). **k** Indel formation at five off-target sites in (**g**). Source data are provided as a Source Data file.

engineering TadA8e to improve editing precision could be used to improve GS-ABE8e performance in future research. The evaluation of base editing also showed no difference in PAM preference between GS-ABE8e and ABE8e (Supplementary Table. 2).

Discussion

We reached the inspirations for REC lobe expansion in Cas9 after systemically analyzing thousands natural sequences and realized artificially enlarging SpCas9. To our knowledge, this is the first report that evolution-inspired engineering of Cas9 domain expansion can improve gene editing performance. Extensive structural study is needed to understand this enlarged Cas9, which may offer novel insights for innovative strategies of optimizing gene editors. Structural growth trajectory of II-C Cas9s was predicted in a recent study[66], which indicates a vast potential for engineering the topological malleability of RNA-guided endonucleases. The ratio of REC lobe size to NUC (nuclease) lobe size may play crucial roles in regulating function of Cas9 or Cas9-based gene editors (Supplementary Fig. 28). Considering the highly conserved bilobed structures of Cas9 orthologs, this largest REC domain could possibly be expanded to other Cas9s. Analysis of structures also shows REC lobe extension parallel with other domain extensions in another RNA-guided-endonuclease family in which Fanzor and Cas12a derived from an ancestor TnpB[67]. Our strategy to create diverse Cas9 variants by enlarging specific domains in the laboratory instead of natural evolution improves the efficacy of the editors. However, we should acknowledge that increasing the size of gene editors could pose challenges for delivery using adeno-associated viruses (AAV). Alternative delivery approaches, such as the split AAV method or LNP-based non-viral systems, could provide potential solutions to this challenge. For example, recent research showed that split inteins can enable AAV delivery of proteins as large as 470 kDa[68]. These developments offer potential solutions for delivering larger gene editors.

Importantly, these crucial insights on REC-expansion may spur further fundamental research to deepen the understanding of RNA-guided-endonuclease topology and stimulate more innovative solutions to overcome the critical challenges associated with gene editors. We believe that this approach is complementary to previous engineering strategies of RNA-guided gene editors, holding great potential advancing safer gene editing for development and optimization of clinical drug.

Methods

Analysis of sequences

Alignment and annotation analyses were performed with Geneious Prime 11.0.18. Filtration of sequences was performed using Microsoft Excel. NCBI blastp (protein-protein BLAST) was used to double check candidate-sequences.

If no other instructions are given, in the text, FnCas9 specifically refers to PDB:5B2O, SpCas9 specifically refers to PDB:4OO8, St1Cas9 specifically refers to PDB:6M0V, SaCas9 specifically refers to PDB:5AXW, NmeCas9 specifically refers to PDB:6JDQ, CjCas9 specifically refers to PDB: 6JOO, OgeuIscB specifically refers to PDB:7UTN.

Construction of plasmids

We used general cloning methods including Gibson assembly with GeneArt Gibson Assembly HiFi Master Mix (Thermo Fisher, A46628) and Quick ligation kit (New England Biolabs, M2200S) with a variety of type IIS restriction enzymes. One Shot™ MAX Efficiency™ DH5α-T1R Competent Cells (Thermo Fisher, 12297016) were used for DNA cloning. The sequences of cloned constructs were confirmed by sanger sequencing following extraction with QIAprep kit (QIAGEN, 27106). Plasmids including vectors purchased from Addgene used in this study can be found in Supplementary Table. 3 to 5. All human genome sites are available in Supplementary Table. 6. Oligonucleotides and fragments used in this research can be found in the Supplementary Table. 7 to 10. EGFP expression plasmids containing amino acid substitutions were generated by standard PCR with Q5 Site-Directed Mutagenesis Kit (New England Biolabs, E0554S). Human-codon-optimized fragments used for REC expansions, oligonucleotides and sgRNA expression plasmids were synthesized by IDT (Integrated DNA Technologies, USA). Unless otherwise indicated, all sgRNAs were designed to target sites containing a 5′ guanine nucleotide. Plasmids encoding GS-Cas9 and GS-ABE8e variants are available from Addgene.

Cell culture and transfection

HEK293T cells (ATCC, CRL-3216) and HEK293-deGFP cells[35] (a gift from Prof. Eben Alsberg, University of Illinois at Chicago) were cultured in DMEM (Sigma, D6429) supplemented with 10% FBS and 1% Gibco™ Penicillin-Streptomycin (10,000 U/mL) at 37 °C and 5% CO_2. Cells were seeded one day prior to transfection in 24-well plates. For screening Cas9 variants, cells were plated at a density of approximately 70,000 cells per well, and cells were transfected with 200 ng of nuclease editor expression plasmid DNA, 200 ng of sgRNA plasmid, and 10 ng of reporter plasmid (pCMV-GFP plasmid). For the pbREADER assay, cells were plated at a density of approximately 70,000 cells per well, and cells were transfected with 40 fmol of base editor expression plasmid DNA, 200 ng of sgRNA plasmid, and 200 ng or 10 ng of reporter plasmid (200 ng of EGFPstop plasmid, 10 ng of pCMV-GFP plasmid). Unless otherwise noted, cells were plated at a density of approximately 50,000 cells per well for in vitro cell genome editing, cells were transfected with 70 fmol of nuclease or 80 fmol base editor expression plasmid DNA and 200 ng of sgRNA plasmid per well. Transfections were performed with Lipofectamine 2000 (Invitrogen, 11668027) and Lipofectamine 3000 (Thermo Fisher, L3000008) according to the manufacturer's recommended protocol in the cpREADER assay and genome editing, respectively. Of note, the ratio of lipofectamine to DNA was set at 2.5:1. NEBioCalculator version 1.15.5 was used to convert dsDNA mass to moles of dsDNA.

The EGFPstop mRNA activation assay was performed according to the following protocol. mRNA carried a stop codon UAG in place of UGG at the 58-position, and mRNA was prepared as described in our previous study[64]. dsDNA templates were prepared in a 50 μl reaction containing 100 ng pMRNA-GFPstop plasmid, 25 μl NEBNext High-Fidelity 2X PCR Master Mix, and 5 μl Tail primer mix (SBI, MR-TAIL-PR). Templates were transcribed using the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB, E2040S) and the standard RNA synthesis protocol in a total volume of 20 μl for 2 h at 37 °C. For unmodified mRNA, 40 mM m7G(5′)ppp(5′)G RNA Cap Structure Analog (NEB, S1404), clean cap GG (Trilink, N-7133-1), and 100 mM each of ATP, UTP, GTP, and CTPs were used. RNA was purified using a Monarch RNA Cleanup Kit (NEB, T2040). HEK293T cells were plated at a density of approximately 70,000 cells per well in 24-well plates. Cells were transfected with 200 ng of ABE8e plasmid or ABE9 plasmid using Lipofectamine 3000. After 6 h, cells were transfected with 400 ng of mRNA using Lipofectamine 2000. Then 48 h, cells were analyzed using a BD Accuri™ C6 Flow Cytometer.

Western blotting assay

HEK293T cells were lysed 3 days after transfection using RIPA buffer complemented with proteinase and phosphatase inhibitors (Pierce, Protease Phosphatase Inhibitor Tablets, Thermo Fisher Scientific: A32959). The total protein concentrations of cell lysate supernatants were quantified using a BCA protein assay kit (Thermo Fisher Scientific). In total, 20 μg of total protein per well was loaded for electrophoresis using a 10-well 4-12% Bis-Tris Gel (Invitrogen, NW04125BOX) and transferred using an XCell II™ Blot Module and PVDF (0.45-μm pore size)). Bolt™ Transfer Buffer (Invitrogen, BT0006) containing 5% methanol was selected as transferring buffer. 25 min and 70 min at 30 V were used for transferring small protein (<50 kDa) and large

protein (>150 kDa), respectively. The gel was stained using Coomassie Blue to confirm transferring efficiency. The membranes were blocked with 5% Non-Fat Dry Milk (AmericanBio, AB10109) for 1 h at room temperature and then divided and processed with different primary antibodies including anti-beta-actin (1:10,000; Abcam, ab49900) and the anti-Flag (1:5000 dilution; Abcam, ab205606) separately overnight at 4 °C. Then, the membranes were incubated with Goat Anti-Rabbit IgG H&L (HRP) (1:10,000 dilution; Abcam, ab205718) for 1 h and visualized using the Bio-Rad imaging system (ChemiDoc™ MP Imaging System). Blot was incubated in chemiluminescent substrate (Thermo Scientific, A38554) for 5 min prior to imaging.

Genomic DNA isolation

Cells were harvested 3 days post transfection. Genomic DNA was extracted from transfected cells using DNeasy Blood & Tissue Kit (Qiagen, 69504) following the manufacturer's protocol. Extracted DNA was normalized to a final concentration of 20 ng per µl with ddH_2O.

Flow cytometry detection

Adherent cells were treated with 100 µl of 0.25% Trypsin-EDTA (Gibco, 25200056) incubated for 5 min at 37 °C to completely detach cells. 400 µl of DMEM was used to stop trypsin digestion. Samples were applied to the BD Accuri™ C6 Flow Cytometer (BD Biosciences) directly, and GFP fluorescence was measured. Data analysis was performed with BD Accuri C6 Software and Microsoft Excel. Prism (GraphPad) was used to generate column graphs and for calculations. EGFP disruption or EGFPstop activation experiments in pbREADER assays were performed as the following protocol. Briefly, transfected cells were analyzed 48 h after transfection for loss or restoration of EGFP fluorescence. The background was determined by gating a negative control transfection.

RNA isolation from mammalian cells

HEK293T cells were transfected with nSpCas9-EGFP, ABE8e-EGFP or GS-ABE8e-EGFP plasmid. 48 h post-transfection, ~300,000 cells were harvested for mRNA isolation. RNA isolation was performed with the RNeasy Plus Mini Kit (QIAGEN, 74134) according to the manufacturer's instructions. In short, RNA isolation began with removal of the culture medium and washing the cells with 1× PBS (Thermo Fisher Scientific). 350 µl of RLT Plus Buffer was added into each well; cells were homogenized by pipetting and transferred into a DNA eliminator column, and the subsequent binding and washing steps for RNA isolation using the RNeasy columns were performed as recommended by the manufacturer. Upon elution of RNA from the RNeasy column with 45 µl of RNase (ribonuclease) free water (QIAGEN), 2 µl of RNase inhibitor (New England Biolabs, M0314S) was added to prevent RNA degradation, and RNA was stored at −80 °C.

cDNA synthesis

cDNA synthesis was performed using ProtoScript II First Strand cDNA Synthesis Kit (New England Biolabs, E6560) with 1 µg of RNA in a total reaction volume of 20 µl. Reactions were incubated for 2 hr at 42 °C. 2 µl cDNA was input into 50 µl NEBNext High-Fidelity 2X PCR Master Mix (New England Biolabs, M0541S) containing specific primers. The purification of target products was performed with GeneJET PCR Purification Kit (Thermo Fisher, K0702).

Preparation of genomic DNA amplicons for deep sequencing

Targeted regions flanking the on-target or off-targeted sites were amplified with specific primers and Q5 Hot Start High-Fidelity 2X Master Mix (NEB, M0494S) under the following thermal cycling conditions: one cycle, 98 °C, 1 min; one cycle, 98 °C, 30 s; 35 cycles, 98 °C, 10 s, 65 °C, 30 s, 72 °C, 15 s; one cycle, 72 °C, 2 min; 4 °C hold. 100 ng of isolated genomic DNA was input into each 50 µl of PCR. PCR products were analyzed on an agarose gel electrophoresis system to verify both size and purity. PCR products were purified with GeneJET PCR Purification Kit (Thermo Fisher, K0702) followed by normalizing to a final concentration of 20 ng per µl with ddH2O.

Analysis of HTS data for targeted amplicon sequencing

Targeted amplicon sequencing was carried out by Genwiz (Azenta, South Plainfield, NJ, US) using the Amplicon-EZ protocol. Batch analysis with the CRISPResso2 pipeline was used for targeted amplicon sequencing. For DNA analysis, a 30-bp window was used to quantify indels around the DNA nick site. Otherwise, the default parameters were used for analysis. The output file "NUCLEOTIDE_PERCENTAGE_TABLE.txt" was imported into Microsoft Excel for quantification of editing frequencies and "Indel_histogram.txt" for quantification of indel frequencies. Indel percentage was re-checked with BEanalyzer if requested. For analysis of RNA amplicon editing, no sgRNA flag was used. Instead, the output file "NUCLEOTIDE_PERCENTAGE_TABLE.txt" was imported into Microsoft Excel for analysis of A-to-G editing rates associated with each sample (inosine in RNA is read as a guanosine by polymerases).

Prism (GraphPad) was used to generate dot plots and bar plots of these data. For instances in the text where means have been calculated across multiple genomic or transcriptomic loci, the SDs reported represent the SD of the mean for all biological replicates.

Amplicon sequencing using Sanger sequencing

Sanger sequencing results of PCR amplicons were analyzed using EditR version 1.0.10 (https://moriaritylab.shinyapps.io/editr_v10/).

Orthogonal R-loop assay

Orthogonal R-loop assays were performed to measure Cas9-independent off-target editing as described previously, with minor modifications. Under standard conditions, 200 ng of base editor plasmid, 300 ng of dSaCas9 plasmid, 100 ng of SpCas9 sgRNA plasmid and 100 ng of SaCas9 sgRNA plasmid were co-transfected into HEK293T cells using 1.5 µl of Lipofectamine 3000. Cells were cultured for 4 days after treatment followed by genomic DNA isolation.

Statistical analysis and graphical illustrations

Curve plotting and statistical analysis were performed using Prism 8 (GraphPad, La Jolla, CA). Data are shown as means ± standard error of the mean for groups of two or more replicates or as individual values with the mean indicated. Graphical illustrations were created using Office PowerPoint.

Statistics and reproducibility

All bar plots display individual biological replicates as dots, as specified in each figure. Detailed statistical information is provided in the figure legends or descriptions. No statistical method was used to predetermine sample size. Sample sizes were based on observed variability across independent experiments and were consistent with standard practices in related research. No data was excluded from the analysis. Investigators were not blinded to group allocation during experiments or outcome assessment. Statistical significance was defined as $P < 0.05$, with "ns" indicating non-significance. *P*-values are provided in the Source Data.

Reporting summary

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

Data availability

High-throughput DNA sequencing FASTQ files generated in this study have been deposited in the National Center of Biotechnology's Information Sequence Read Archive under BioProject "PRJNA1121945 and PRJNA1121941". Amino acid and DNA sequences of RECs designed in

this work are listed in the Supplementary Information. Source data are provided with this paper. The published structure of FnCas9 (5B2O), SpCas9 (4OO8), SaCas9 (5AXW), St1Cas9 (6M0V), NmeCas9 (6JDQ), CjCas9 (6JOO), OgeuIscB (7UTN), and ABE8e (6VPC) are available in the Protein Data Bank. Source data are provided with this paper.

References

1. Altae-Tran, H. et al. The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).
2. Altae-Tran, H. et al. Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science* **382**, eadi1910 (2023).
3. Oakes, B. L. et al. CRISPR-Cas9 circular permutants as programmable scaffolds for genome modification. *Cell* **176**, 254–267.e216 (2019).
4. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
5. Kapitonov, V. V., Makarova, K. S., Koonin, E. V. & Zhulin, I. B. ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. *J. Bacteriol.* **198**, 797–807 (2016).
6. Oakes, B. L. et al. Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nat. Biotechnol.* **34**, 646–651 (2016).
7. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
8. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
9. Chen, J. S. et al. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
10. Casini, A. et al. A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).
11. Lee, J. K. et al. Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat. Commun.* **9**, 3048 (2018).
12. Kim, Y.-h. et al. Sniper2L is a high-fidelity Cas9 variant with high activity. *Nat. Chem. Biol.* **19**, 972–980 (2023).
13. Wang, J. Y. & Doudna, J. A. CRISPR technology: a decade of genome editing is only the beginning. *Science* **379**, eadd8643 (2023).
14. Hino, T. et al. An AsCas12f-based compact genome-editing tool derived by deep mutational scanning and structural analysis. *Cell* **186**, 4920–4935.e4923 (2023).
15. Madigan, V., Zhang, F. & Dahlman, J. E. Drug delivery systems for CRISPR-based genome editors. *Nat. Rev. Drug Discov.* **22**, 875–894 (2023).
16. Gillmore, J. D. et al. CRISPR-Cas9 in vivo gene editing for transthyretin amyloidosis. *N. Engl. J. Med* **385**, 493–502 (2021).
17. Musunuru, K. et al. In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates. *Nature* **593**, 429–434 (2021).
18. Gao, S. & Xu, Q. New ionizable lipids reduce the lipid-to-mRNA ratio for base editing. *Natl Sci. Rev.* **11**, nwae224 (2024).
19. Nishimasu, H. et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
20. Nishimasu, H. et al. Crystal structure of Staphylococcus aureus Cas9. *Cell* **162**, 1113–1126 (2015).
21. Hirano, H. et al. Structure and engineering of Francisella novicida Cas9. *Cell* **164**, 950–961 (2016).
22. Jiang, F. et al. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**, 867–871 (2016).
23. Yamada, M. et al. Crystal structure of the minimal Cas9 from Campylobacter jejuni reveals the molecular diversity in the CRISPR-Cas9 systems. *Mol. Cell* **65**, 1109–1121.e1103 (2017).
24. Sun, W. et al. Structures of Neisseria meningitidis Cas9 complexes in catalytically poised and anti-CRISPR-inhibited states. *Mol. Cell* **76**, 938–952.e935 (2019).
25. Schuler, G., Hu, C. & Ke, A. Structural basis for RNA-guided DNA cleavage by IscB-ωRNA and mechanistic comparison with Cas9. *Science* **376**, 1476–1481 (2022).
26. Seo, S.-Y. et al. Massively parallel evaluation and computational prediction of the activities and specificities of 17 small Cas9s. *Nat. Methods* **20**, 999–1009 (2023).
27. Shams, A. et al. Comprehensive deletion landscape of CRISPR-Cas9 identifies minimal RNA-guided DNA-binding modules. *Nat. Commun.* **12**, 5664 (2021).
28. Pacesa, M. et al. R-loop formation and conformational activation mechanisms of Cas9. *Nature* **609**, 191–196 (2022).
29. Fonfara, I. et al. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res* **42**, 2577–2590 (2014).
30. Elowitz, M. & Lim, W. A. Build life to understand it. *Nature* **468**, 889–890 (2010).
31. Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A. & Clarke, J. The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **8**, 319–330 (2007).
32. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300**, 1701–1703 (2003).
33. Zhang, Y. et al. Catalytic-state structure and engineering of Streptococcus thermophilus Cas9. *Nat. Catal.* **3**, 813–823 (2020).
34. Müller, M. et al. Streptococcus thermophilus CRISPR-Cas9 systems enable specific editing of the human genome. *Mol. Ther.* **24**, 636–644 (2016).
35. Andreatta, C. et al. Use of short-lived green fluorescent protein for the detection of proteasome inhibition. *BioTechniques* **30**, 656–660 (2001).
36. Ma, E., Harrington, L. B., O'Connell, M. R., Zhou, K. & Doudna, J. A. Single-stranded DNA cleavage by divergent CRISPR-Cas9 enzymes. *Mol. Cell* **60**, 398–407 (2015).
37. Yourik, P., Fuchs, R. T., Mabuchi, M., Curcuru, J. L. & Robb, G. B. Staphylococcus aureus Cas9 is a multiple-turnover enzyme. *RNA* **25**, 35–44 (2019).
38. Nguyen Tran, M. T. et al. Engineering domain-inlaid SaCas9 adenine base editors with reduced RNA off-targets and increased on-target DNA editing. *Nat. Commun.* **11**, 4871 (2020).
39. Acharya, S. et al. Francisella novicida Cas9 interrogates genomic DNA with very high specificity and can be used for mammalian genome editing. *Proc. Natl Acad. Sci. USA* **116**, 20959–20968 (2019).
40. Jang, H.-K. et al. High-purity production and precise editing of DNA base editing ribonucleoproteins. *Sci. Adv.* **7**, eabg2661 (2021).
41. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* **38**, 1328–1336 (2020).
42. Schmid-Burgk, J. L. et al. Highly parallel profiling of Cas9 variant specificity. *Mol. Cell* **78**, 794–800.e798 (2020).
43. Lapinaite, A. et al. DNA capture by a CRISPR-Cas9–guided adenine base editor. *Science* **369**, 566–571 (2020).
44. Richter, M. F. et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat. Biotechnol.* **38**, 883–891 (2020).
45. Zeng, H. et al. A split and inducible adenine base editor for precise in vivo base editing. *Nat. Commun.* **14**, 5573 (2023).
46. Liu, Y. et al. A Cas-embedding strategy for minimizing off-target effects of DNA base editors. *Nat. Commun.* **11**, 6073 (2020).
47. Villiger, L. et al. Replacing the SpCas9 HNH domain by deaminases generates compact base editors with an alternative targeting scope. *Mol. Ther. Nucleic Acids* **26**, 502–510 (2021).
48. Chen, L. et al. Engineering a precise adenine base editor with minimal bystander editing. *Nat. Chem. Biol.* **19**, 101–110 (2023).

49. Neugebauer, M. E. et al. Evolution of an adenine base editor into a small, efficient cytosine base editor with low off-target activity. *Nat. Biotechnol.* **41**, 673–685 (2023).
50. Aurora, R., Creamer, T. P., Srinivasan, R. & Rose, G. D. Local interactions in protein folding: lessons from the α-Helix *. *J. Biol. Chem.* **272**, 1413–1416 (1997).
51. Rees, H. A., Wilson, C., Doman, J. L. & Liu, D. R. Analysis and minimization of cellular RNA editing by DNA adenine base editors. *Sci. Adv.* **5**, eaax5717 (2019).
52. Rodriguez, C., Tompkin, J., Hazel, J. & Foster, P. L. Induction of a DNA nickase in the presence of its target site stimulates adaptive mutation in Escherichia coli. *J. Bacteriol.* **184**, 5599–5608 (2002).
53. Xu, S.-y & Gupta, Y. K. Natural zinc ribbon HNH endonucleases and engineered zinc finger nicking endonuclease. *Nucleic Acids Res* **41**, 378–390 (2013).
54. Doman, J. L. Raguram, A., Newby, G. A. & Liu, D. R. Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. *Nat. Biotechnol.* **38**, 620–628 (2020).
55. Hendel, A. et al. Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.* **33**, 985–989 (2015).
56. Lei, Z. et al. Detect-seq reveals out-of-protospacer editing and target-strand editing by cytosine base editors. *Nat. Methods* **18**, 643–651 (2021).
57. Ranzau B. L., Rallapalli K. L., Evanoff M., Paesani F., Komor A. C. The wild-type tRNA adenosine deaminase enzyme TadA is capable of sequence-specific DNA base editing. *ChemBioChem* **n/a**, e202200788 (2023).
58. Yang, H. & Patel, D. J. Inhibition mechanism of an Anti-CRISPR suppressor AcrIIA4 targeting SpyCas9. *Mol. Cell* **67**, 117–127.e115 (2017).
59. Alves C. R. R., et al. Optimization of base editors for the functional correction of SMN2 as a treatment for spinal muscular atrophy. *Nat. Biomed. Eng.* **8**, 118-131 (2023).
60. Tálas, A. et al. BEAR reveals that increased fidelity variants can successfully reduce the mismatch tolerance of adenine but not cytosine base editors. *Nat. Commun.* **12**, 6353 (2021).
61. Sretenovic, S. et al. Genome- and transcriptome-wide off-target analyses of a high-efficiency adenine base editor in tomato. *Plant Physiol.* **193**, 291–303 (2023).
62. Sangree, A. K. et al. Benchmarking of SpCas9 variants enables deeper base editor screens of BRCA1 and BCL2. *Nat. Commun.* **13**, 1318 (2022).
63. Grunewald, J. et al. A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing. *Nat. Biotechnol.* **38**, 861–864 (2020).
64. Gao, S. et al. Harnessing non-Watson–Crick's base pairing to enhance CRISPR effectors cleavage activities and enable gene editing in mammalian cells. *Proc. Natl Acad. Sci. USA* **121**, e2308415120 (2024).
65. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. *Nat. Biotechnol.* **41**, 663–672 (2023).
66. Zhang S. et al. Pro-CRISPR PcrIIC1-associated Cas9 system for enhanced bacterial immunity. *Nature*, **630**, 484–492 (2024).
67. Saito M. et al. Fanzor is a eukaryotic programmable RNA-guided endonuclease. *Nature*, **620**,660–668 (2023).
68. Zhou, Y., Zhang, C., Xiao, W., Herzog, R. W. & Han, R. Systemic delivery of full-length dystrophin in Duchenne muscular dystrophy mice. *Nat. Commun.* **15**, 6141 (2024).

Acknowledgements

This work was supported by the US National Institutes of Health (UG3TR002636) and Hopewell Therapeutics Inc. We thank Prof. Feng Zhang and Blake Lash from Broad Institute of MIT and Harvard for valuable discussions. We thank Prof. Eben Alsberg and Dr. Cong Truc Huynh from University of Illinois at Chicago's Department of Biomedical Engineering for the gift of HEK293-deGFP cells. We thank Dr. Xiaoli Zhang from Tufts University's Department of Biomedical Engineering for assistance with RNA isolation and analysis. We thank Prof. Luca Pinello from Harvard Medical School and Edilytics Inc's CEO Cole Lyman for help on CRISPResso2 troubleshooting. We thank Dr. Menglin Wang from Massachusetts Institute of Technology's Whitehead Institute for Biomedical Research for discussions on protein analysis.

Author contributions

Conceptualization: S.G. and Q.X. Funding acquisition: Q. X. Experiments design: S.G. Experiments execution: S.G., B.W., D.W., M.C., H.G., Z.Y., and C.X. Data analysis: S.G. and B.W. Writing and editing: S. G., D.W., L.P., and Q. X.

Competing interests

Tufts University has submitted a patent application on behalf of S.G. and Q.X. on the gene editors developed in this work (application number 63/558,998). The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57154-5.

Correspondence and requests for materials should be addressed to Qiaobing Xu.

Peer review information *Nature Communications* thanks Zhiquan Liu, Shaohua Yao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025