# Cooperativity between Cas9 and AID establishes broad and diversifying mutational footprints in base editors

Kiara N. Berríos[1], Aleksia Barka[2], Jasleen Gill[2], Niklaus H. Evitt[3], Kiran S. Gajula[2], Junwei Shi[4*] and Rahul M. Kohli[2*]

[1]Graduate Group in Biochemistry and Molecular Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA.
[2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA.
[3]Graduate Group in Cell and Molecular Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA.
[4]Department of Cancer Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA.

* Correspondence to Junwei Shi and Rahul M. Kohli: jushi@upenn.edu and rkohli@pennmedicine.upenn.edu

## SUMMARY

**The partnership of DNA deaminase enzymes with CRISPR-Cas nucleases is now a well-established method to enable targeted genomic base editing. However, an understanding of how Cas9 and DNA deaminases collaborate to shape base editor (BE) outcomes has been lacking. Here, we employ a range of hyperactive AID base editors (hBEs) to deduce a novel mechanistic model that reveals multiple layers of previously underappreciated cooperativity in BE steps. These include: (1) Cas9 binding can potentially expose both DNA strands for 'capture' by the deaminase, a feature enhanced by guide mismatches; (2) after strand capture, intrinsic deaminase activity can tune window size and base editing activity; and (3) non-canonical base edits can be further elicited by modulating Cas9 nicking activity. Leveraging insights from our mechanistic model, we create novel hBEs that can remarkably generate simultaneous C>T and G>A transitions over >65 bp with the potential for gene diversification.**

## INTRODUCTION

Cytosine base editors (BEs) combine the powerful DNA deamination activity of AID/APOBEC family enzymes with CRISPR-Cas targeting in order to introduce point mutations at a specified locus.[1] In the first generation of base editors (BE1), rat APOBEC1 (rA1) was fused to a catalytically dead *S. pyogenes* Cas9 (dCas9) and targeted using a single-guide RNA (sgRNA).[2] Upon dCas9-sgRNA binding to the target strand (TS), R-loop formation exposes the non-target strand (NTS) resulting in single-stranded DNA (ssDNA) that is then 'captured' by the DNA deaminase. The NTS can then be deaminated in this exposed 'editing window', typically 12-16 base pairs (bp) away from the protospacer adjacent motif (PAM), and the resulting C/G>T/A mutations are locked in after replication over the lesion.

Since BE1, several strategies have been employed to increase precise editing or alternatively to generate diversifying mutations at the target site.[2-9] Major improvements to BEs that enhance efficiencies include: (1) the incorporation of UGI – a small phage-derived inhibitor of uracil DNA glycosylase – to suppress base excision repair in BE2 editors,[2] (2) the replacement of dCas9 with a TS-nickase (Cas9-D10A, nCas9) to promote replication over NTS mutations in BE3 editors,[2] and (3) mammalian codon optimization along with the addition of a second UGI in BE4max editors.[10,11] Novel BEs with diverse features have also been developed by using deaminases beyond rA1, including human AID or APOBEC3 enzymes, as well as engineered TadA variants for adenine base editors (ABEs). [4,12-14] Additionally, Cas9 engineering has been employed to alter outcomes, with PAM-modified or circularly permuted Cas9 variants increasing the targeting scope of base editors.[15,16]

Given the transformative potential of BEs, intense efforts have focused on the development of improved tools. While these remarkable advances have now yielded a broad and powerful toolbox for genome engineering,

fundamental mechanistic questions remain unanswered. In particular, the critical question of how Cas9 and DNA deaminases coordinate to promote BE outcomes has yet to be fully explored. The current model for BE-mediated mutagenesis focuses on mutation of the NTS and involves multiple steps: Cas9 binding and unwinding of the DNA duplex, AID/APOBEC capture of the ssDNA bubble and enzymatic deamination, along with TS nicking and replication over the uracil lesion (**Fig. 1a**). When considering this stepwise mechanism, illuminating the interplay of Cas9 engagement and DNA deaminase action in each step could facilitate the continued optimization and expansion of BEs and support the development of the next generation of genome engineering tools.

As noted, the existing toolbox for BEs has utilized diverse members of the AID/APOBEC family, each with distinctive sequence preferences, intrinsic deamination rates, cellular localization, and modes of action. As these features are retained when DNA deaminases are incorporated into BEs, it has been challenging to decipher the impact of *single* attributes on the BE mechanism. By keeping the BE scaffold the same and manipulating specific attributes, the underlying reaction mechanisms could be better elucidated. In this regard, AID provides several distinctive features that make it an ideal candidate for systematic variation. AID naturally functions to drive somatic hypermutation (SHM) in B-cells where it generates clustered mutations on the immunoglobulin (Ig) locus by acting on either strand in a processive fashion.[17,18] In addition to being processive, AID activity can be tuned as its deamination activity has been highly constrained,[19] as evidenced by the isolation of various hyperactivating mutations that shift the balance towards increased genomic instability in B-cells.[20,21] Similar to the opportunity offered by systematically altering AID, perturbing Cas9 cleavage activity can offer a means to alter downstream replication and repair, and analyzing BE off-targets can provide a means to assess the impact of Cas9 engagement. Thus, AID-focused BEs offer an ideal model system for understanding the role of each step in the BE mechanism.

In this study, we aimed to create a complete mechanistic model for how collaboration between Cas9 and AID establish BE outcomes. To this end, we derive a range of hyperactive AID base editors (hBEs), and reveal that BE activity and editing windows can be directly tuned by altering the intrinsic activity of the associated DNA deaminase. By further perturbing Cas9 engagement and downstream nicking and repair, hBEs provide unexpected new insights, revealing that both NTS and TS can be captured by the enzyme for subsequent deamination. Both mismatched sgRNAs and Cas9 nicking variants can increase unexpected editing at the TS, resulting in a striking increase in G>A mutations neighboring C>T changes. Beyond advancing a complete model for the impact of Cas9 binding, deaminase strand capture, deamination, and DNA repair in BEs, our work yields a simplified diversifying BE with a distinctive and novel capability of efficiently generating simultaneous C>T and G>A transitions over a >65 bp window. Our systematic approach shows how BE mutational landscapes are cooperatively established by Cas9 and AID and how modifying their activities can uncover new gene editing outcomes for diverse biotechnological applications.

## RESULTS

**Modulating intrinsic AID activity tunes BE activity**. To first generate a potential range of hyperactive AID variants with tunable activity, we permuted mutations previously shown to enhance AID activity (**Fig. S1a**). One cluster of mutations [C1] (AID-C1, **Fig. 1b**), developed during a comprehensive structure-function analysis of AID, involved multiple changes in the ssDNA-binding loop, with R119G as a dominant activating mutation.[21] A second cluster [C2] (AID-C2, **Fig. 1b**), uncovered using an iterative selection strategy, included T82I as a dominant mutation.[20] Prior work has also shown that a naturally occurring variant lacking the last exon (denoted as AID*, **Fig. 1b**) also shows enhanced activity over wild-type AID (AID-WT).[22]

We generated a total of eight AID variants with candidate hyperactivating mutations. To measure deaminase activity, we employed the well-established rifampin mutagenesis assay,[21,23] whereby variants are overexpressed in *E. coli* and the frequency of acquired resistance to rifampin ($Rif^R$) provides a reliable surrogate of deaminase activity (**Fig. 1c**). In this assay, AID-WT increases $Rif^R$ 11-fold relative to a catalytically-inactive AID-E58A, while AID* shows a further 4-fold increase over AID-WT. Adding single activating mutations (R119G or T82I) to AID* leads to a further increase, while we observed a 40-fold and 75-fold increase in $Rif^R$ for AID-C1* and AID-C2*, respectively, relative to AID-WT. Strikingly, combining all mutations in AID-C12* results in a 285-fold increase in $Rif^R$ (**Fig. 1c**). The synergy between these independently derived hyperactivating mutations further supports the claim that AID has been naturally attenuated, likely to modulate risks posed to the genome, and that AID can be further engineered to improve its catalytic activity.

The development of a spectrum of AID variants positioned us to systematically explore the relationship between intrinsic deaminase activity and base editor outcomes. To this end, we incorporated the AID variants into BE3 constructs, creating a range of potentially hyperactive AID base editors (hBEs). To measure editing activity, we employed a reporter cell line with a single copy of destabilized GFP (*d2gfp*) integrated into HEK293T.[24] Editing can be tracked by flow cytometry using an sgRNA that targets *d2gfp* to either generate indels with WT Cas9 or a nonsense Q158* mutation with base editing (**Fig. 1d**). Using this assay, WT Cas9 editing is efficient, with 85±4% of GFP+ reporter cells converted to $GFP^{off}$. With the same sgRNA, minimal GFP inactivation was observed with AID-WT-BE3. In line with our hypothesis, however, GFP inactivation significantly increases with hBEs (**Fig. 1d, Extended Data Fig. 1b**). Editing by AID*-BE3 results in 22±7% $GFP^{off}$, with further enhancement by with R119G and T82I mutants. AID-C1*-BE3 and AID-C2*-BE3 showed 26±4.8% and 35±8.6% $GFP^{off}$, respectively, while the full suite of synergistic hyperactivating mutations in AID-C12*-BE3 yields 53±2% $GFP^{off}$. To integrate across the series, we plotted $Rif^R$ frequency, a measure of intrinsic deaminase activity, against $GFP^{off}$ frequency, a measure of BE activity (**Fig. 1e**). An overall strong and positive correlation ($r = 0.89$) suggests that the intrinsic deaminase activity can be rate-limiting in the complex multi-step BE reaction.

**Deaminase activity shapes editing window.** To better understand the mutational footprints of hBEs, we next deep-sequenced the *d2gfp* locus for five constructs spanning a range of intrinsic deaminase activities. With hBE constructs, target cytosine conversion (C > T/G/A/indels) increased relative to AID-WT-BE3 but appeared comparable across constructs (**Fig. Extended Data Fig. 1c**), indicating that increased *d2GFP* inactivation did not exclusively come from nonsense mutation at the Q158 codon. When broadening analysis, we observed a striking pattern with multiple mutations across the locus, including C>T edits on both DNA strands (**Fig. 2a – Left).** To help organize observations, we define NTS editing as C>T editing on the top strand (schematic in **Fig. 1a**), and TS editing as C>T editing on the bottom strand, which is read as G>A mutations on the top strand. With hBEs, we observed both NTS (C>T) and TS (G>A) editing events, including outside the traditional BE window. At the level of individual DNA strands mutations also appear clustered tracking with increasing deaminase activity.

To dissect the observed trends, we assessed the mutational load of the *d2GFP* locus after hBE treatment by quantifying the average number of C>T and G>A edits per mutated read (**Fig. 2a – Right**). Cells treated with AID-WT-BE3 and AID*-BE3 harbored similar mutational loads, 1.5±0.2 and 1.6±0.1 mutations per read (mutations/read) respectively, however, AID*-BE3 reported higher activity on all sites in the *d2gfp* locus. AID-C1*-BE3 and AID-C2*-BE3, which have comparable intrinsic deaminase rates, both generate 2.4±0.4 mutations/read while AID-C12*-BE3 was able to generate on average 4.3±1.1 mutations/read. Notably, we observed reads with as many as 15 combined C>T and G>A mutations/read with AID-C12*-BE3, a feature suggestive of its capacity to be a powerful diversifying base editor. When the average mutations/read was plotted against $Rif^R$ rates, there was a strong correlation (r = 0.98) between intrinsic DNA deaminase activity and the overall mutational load (**Fig. 2b**). To explore the generality of this result, we extended our analysis to a second genomic locus, *TDG,* and replicated our key findings, with tunable mutation loads and a broad editing footprint as a function of intrinsic deaminase activity (**Extended Data Fig. 2**). While we observed that increased efficiency broadens the editing window, our result provides a strong corollary to prior studies where stunting deaminase efficiency narrowed the editing window and increased precision.[13] Our results also highlight that increasing AID catalytic activity not only increases BE activity, but also elicits the native processivity of AID, simulating its behavior on the Ig locus at sites of our preference.

**Cas9 engagement impacts DNA deaminase strand capture.** There are multiple hypotheses addressing how AID targets both Ig strands in B-cells,[19] including antisense transcription, DNA topology, and R-loop collapse by endogenous RNAse H digestion, all of which would expose both strands as ssDNA. However, how Cas9 binding could create the right environment for AID to access either the NTS or TS has not yet been studied. Given clear signatures of both NTS (C>T) and TS (G>A) editing in our constructs, we next aimed to explore the DNA deaminase 'strand capture' step by perturbing Cas9 engagement. We designed our experiment around a method

known as the R-loop assay, which allowed us to probe DNA deaminase activity at three sites with varying Cas9 engagement modes simultaneously.[25] In this assay, HEK293T cells are co-transfected with three plasmids encoding the hBE construct, an *EMX1*-targeting sgRNA, and a catalytically-dead *S. aureus* Cas9 (dSaCas9) together with its sgRNA targeting a totally unrelated genomic locus. Using this system, the impact of Cas9 engagement on editing can be studied at (1) the on-target site (*EMX1*) with matched sgRNA, (2) known off-target sites (*EMX1*-OT1/2) where sgRNA is mismatched, as it is not fully complementary to the off-target sequences, and (3) the dSaCas9 target site where an R-loop is generated *in trans* and independent of the SpCas9-based BE (**Fig. 2c – Top**).

As anticipated based on *d2GFP* and *TDG* results, while editing at the target C of the on-target *EMX1* locus plateaued (**Fig. 2c - Bottom**), we observed a consistent increase in the average total C>T and G>A mutations per read across hBE constructs **(Extended Data Fig. 3)**. By contrast, at *EMX1*-OT1/2, AID-WT-BE3 showed near background levels of editing while hBE constructs show a consistent increase as a function of deaminase hyperactivation (**Fig. 2c - Bottom, S4**). The altered pattern was even more striking when analyzing NTS (C>T) versus TS (G>A) mutations. At sgRNA-dependent off-target sites, unlike at the on-target site, for AID-C12*-BE3, TS (G>A) mutations were equally as prevalent as NTS (C>T) mutations **(Fig. 2d, Extended Data Fig. 4)**, raising the possibility that sgRNA mismatches could permit AID capture and deamination of the TS. The R-loop generated *in trans* by dSaCas9 provided yet another stark contrast. While increased intrinsic deaminase activity still correlated with NTS mutations, TS mutations were nearly undetectable. These results suggest a high degree of cooperativity within BE3, and a clear and novel finding regarding the impact of the Cas9-sgRNA complex on deaminase activity. When the two DNA binding domains, Cas9 and AID, are acting *in cis*, AID can access not only the NTS, but also the TS, and this activity increases significantly with sgRNA mismatches. However, when the DNA deaminase acts *in trans* at an R loop generated by dSaCas9 with a perfectly matched sgRNA, the TS appears inaccessible.

**Cas9 nuclease activity shapes the landscape of NTS vs. TS mutations.** Intrigued that hBEs have broadened editing windows and generate TS mutations, we first aimed to increase hBE activity by moving AID-C12* from the BE3 into the BE4max scaffold.[10] As predicted with this change, when targeting the *d2GFP* locus, AID-C12*-BE4max led to a 1.8-fold increase in target base C>T conversion and a 2.6-fold decrease in deletions when compared to AID-C12*-BE3 (**Extended Data Fig. 5**).

Given the improved performance, we next characterized AID-C12*-BE4max across a broader array of genomic sites. We targeted seven independent genomic loci (**Supplementary Table 1**), allowing us to derive an aggregate profile of C base editing on either strand across an 80 bp region flanking the target site (**Fig. 3a**). Similar to traditional BEs, in these aggregate profiles maximal NTS (C>T) editing by hBEs occurs inside the protospacer 10-20 bp away from the PAM, reaching a maximum of ~25-38% across the seven loci (**Fig. 3a**). Upstream of the

maximal editing window, editing tapers away but can be detected up to 80 bp away from the PAM sequence. However, downstream of the maximal editing window, editing quickly diminishes with mutations appearing undetectable further downstream of the PAM. This observation reveals that Cas9-binding to the PAM obstructs bidirectional AID processivity, skewing mutations to the region upstream of the PAM only.

For G>A transitions, which occur upon AID engagement with the TS, we observe a different pattern. Maximal editing occurs outside the protospacer, peaking ~21-28 bp away from the PAM and reaching ~15% activity on average across the seven loci (**Fig. 3a**). This finding suggests that sgRNA-engagement effectively blocks AID access to the TS in the protospacer region. G>A edits also taper away, but can be detected at >1% as far as 50 bp upstream from the PAM.

To better understand the mechanistic basis for TS mutations, we next analyzed the pattern of mutations in individual sequencing reads (**Fig. 3b**, left most column 'n' for nCas9**)**. Interestingly, the majority of C>T transitions occurred alone. In contrast, G>A transitions were almost exclusively observed in reads that also harbored C>T transitions, and not in isolation. We hypothesized that NTS (C>T) mutations could happen preferentially during a BE 'first encounter'. After DNA replication locks in these C>T mutations, a 'second encounter' with a BE can occur. The presence of mismatches in the sgRNA-binding site arising from the first encounter could then increase access to the TS for editing, which would result in both C>T and G>A transitions on the same read. To further substantiate this model, we looked at the sgRNA-dependent off-target effects at both *EMX1* and *FANCF* off-target sites (**Fig. 3c**). These sites are marked by multiple PAM-distal sgRNA mismatches that would be present in the first encounter. In line with our hypothesis, we observe an increase in G>A mutations overall. More specifically, however, strand-specific analysis revealed a marked increase in G>A only mutations, indicating that in the presence of mismatches, TS editing alone is increased and can happen in a first encounter with AID-C12*-BE4max **(Fig. 3d)**.

**Manipulating Cas9 cleavage activity skews BE outcomes and yields a proficient diversifying editor.** In BE4max, nCas9 nicking of the TS increases the probability of locking in NTS mutations. As such, nick-associated repair could obscure the extent of NTS versus TS editing. To further explore how Cas9 could cooperatively alter base editing outcomes, we next exchanged nCas9 in the AID-C12*-BE4max construct with either dCas9 or Cas9-H840A (n'Cas9), a variant that nicks the NTS rather than the TS, which could increase the probability of locking in TS mutations. To construct an overall targeting profile, we employed these variants to edit the same seven genomic loci as with AID-C12*-BE4max.

As anticipated when losing TS nicking with the dCas9 construct, we observe a decrease in overall C>T editing activity, with a range of ~11-20% maximal editing across the seven loci (**Fig. 3e).** By contrast, we observed an increase in NTS (G>A) transitions, which localizes to the same window (~21-28 bp upstream of PAM) as observed with the nCas9 construct. Interestingly, however, strand-specific analysis, demonstrated that

G>A edits were now readily detectable by themselves in the absence of C>T edits (**Fig. 3b**, middle column 'd' for dCas9). These results support a model where capture and mutation of the TS can occur in normal BE4max constructs in the first encounter, but such mutations can be reversed by the repair initiated at the nick.

Based on the established model, we anticipated that AID-C12*-n'Cas9-BE4max might be a particularly powerful diversifying base editor. Analyzing across the seven target loci, we find that G>A edits increase yet further relative to the dCas9 construct, reaching levels that rival those of C>T editing (**Fig. 3f**). Strand-specific analysis revealed a nearly equal prevalence of sequences with C>T alone, G>A alone or dual C>T and G>A edits, with detectable mutations at a frequency of >1% extending over a 65 bp window (**Fig. 3b**, right most column for n'Cas9). Thus, AID-C12*-n'Cas9-BE4max appears to offer the broadest diversifying window to date for a DNA deaminase-linked BE construct.

Returning to our model, we also examined editing at the *EMX1* and *FANCF* off-target sites **(Fig. 3d)**. Consistent with the model, the dCas9 and n'Cas9 editors show a pattern where G>A only transitions are relatively increased due to the sgRNA mismatches further facilitating TS editing in the first encounter (**Fig. 3g-h**). Hence, our results demonstrate that modulating Cas9 catalytic activity in concert with hyperactivation of AID can mimic somatic hypermutation across highly diverse loci with processive, targeted diversification across strands.

## DISCUSSION

In this report, we substantiate a complete mechanistic model for the BE mechanism that accounts for Cas9 engagement, strand capture by AID, deamination, and DNA repair **(Fig. 4).** Specifically, our study reveals several previously unappreciated means whereby Cas9 and AID collaborate to promote BE outcomes.

First, we show that strand capture by the DNA deaminase is impacted by the nature of the R-loop generated by Cas9. At on-target sites, binding of a matched sgRNA to the TS results in preferential exposure of the NTS, leading to its capture by the deaminase and subsequent generation of C>T mutations. By contrast, at off-target sites, in the presence of sgRNA-TS mismatches, the TS can more readily be accessed and deaminated, resulting in an increase of G>A mutations (reading from the perspective of the NTS). The lack of G>A only mutations detected for AID-C12*-BE4max during a first BE encounter at on-target sites provides orthogonal evidence to support this claim. After the initial introduction of NTS (C>T) mutations, sgRNA mismatches are created, increasing TS capture in a second encounter, and leading to detectable G>A mutations in the presence of the already established C>T mutations.

Second, we show that once the strand has been captured, the overall base editing activity and the editing window can be tuned as a function of the intrinsic activity of the DNA deaminase. By using the same AID scaffold and variants spanning several orders of magnitude in $Rif^R$ activity, we show that intrinsic deaminase activity can govern the overall BE reaction. As activity is further increased, the intrinsic properties of DNA deaminases,

specifically AID's processivity, are evident with strand-specific clustered mutations that can extend far outside of the canonical editing window opened by Cas9.

Third, as with strand capture, we show that there is cooperativity between Cas9 and AID in establishing the mutational footprint. On the NTS, the PAM binding site serves as a barrier preventing downstream mutagenesis. By contrast on the TS, matched sgRNA binding largely blocks mutations on the protospacer, leading to distinctive maximal peaks for C>T versus G>A mutations.

Fourth, we demonstrate that in concert with the DNA deaminase, Cas9's catalytic activity further contributes to editing outcomes. Prior work has shown that moving from dCas9 in BE1 to nCas9 in BE3 leads to higher C>T editing activity.[2] Here, by changing nCas9 to dCas9 in AID-C12*-BE4max we also show that nicking of the TS results in the suppression of rare TS (G>A) editing events. Changing to n'Cas9 results in an even higher degree of TS mutations alone, although the fact that the nick is positioned downstream of the mutations likely limits the ability of DNA replication to lock in mutations as readily.

We conclude by highlighting how dissecting the molecular mechanisms of individual steps involved in a complex multi-step BE reaction offers an effective means to develop new tools. In our case, by hyperactivating AID, altering Cas9 activity via nickase manipulation and studying different modes of Cas9 engagement at off-target sites, we generated a novel diversifying base editor (AID-C12*-n'Cas9-BE4max) that produces C>T only, G>A only, and concurrent C>T and G>A mutations with nearly equal efficiency over a window of more than 65 bp. These features rival alternative approaches that have been taken to achieve diversification, such as dual ABE/CBE strategies, recruitment of multiple deaminases to the same site, or targeted action of error-prone polymerases.[26-29] Taken together, our study highlights how mechanistic inquiries can continue to drive the discovery and expansion of the genome editing toolbox.

## ACKNOWLEDGEMENTS

This work was in part supported by the Penn Center for Genomic Integrity (PCGI) and the US National Institutes of Health (NIH) through R01-GM138908 and R01-HG010646 (to R.M.K.). K.N.B is an NSF Graduate Research Fellow. N.H.E. was supported by NIH T32-GM007170 and F30-HG011578.

## AUTHOR CONTRIBUTIONS

K.N.B., J.S., and R.M.K. conceived of the approach and designed the research. K.N.B., A.B., and K.S.G. performed experiments and analyzed data. K.N.B., A.B., J.G., and N.H.E. performed computational and statistical analysis. The manuscript was drafted by K.N.B., revised by J.S., and R.M.K., with added input and approval from all authors.

## DECLARATION OF INTERESTS

R.M.K. is on the Scientific Advisory Board for Life Edit, Inc. The University of Pennsylvania has filed a patent on aspects of this work.

# MAIN FIGURES

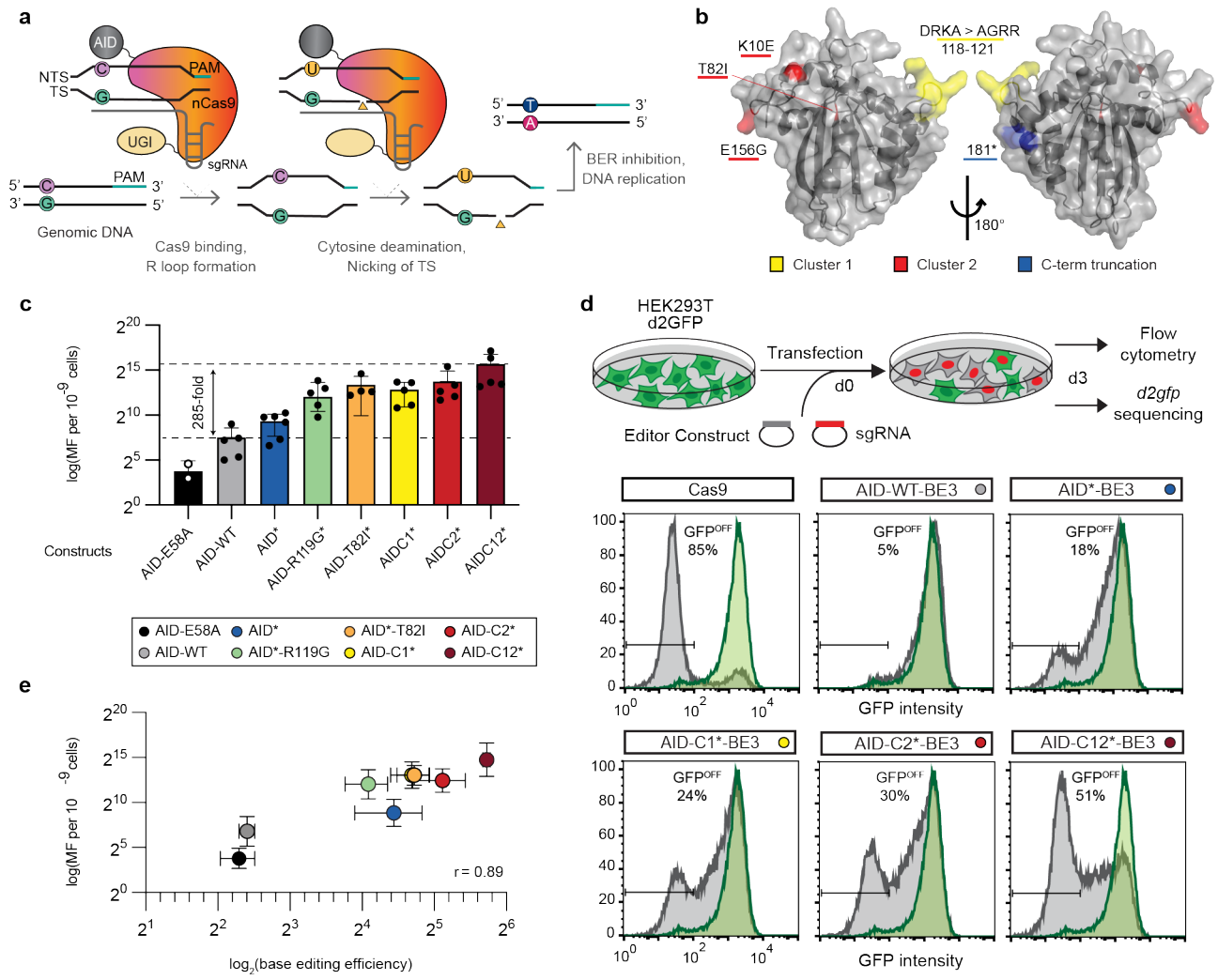



**Figure 1: Hyperactive AID deaminases increase base editing activity. (a)** In base editing, sgRNA-bound nCas9 binds the target strand (TS) forming an R loop exposing ssDNA on the non-target strand (NTS). The DNA deaminase enzyme then converts C>U bases in the exposed ssDNA. TS nicking by nCas9, inhibition of base excision repair (BER) pathway, and replication over the uracil locks in C/T to G/A mutations in genomic DNA. **(b)** Schematic of the AID structure[30] (PDB 5w0z) showing clusters of hyperactivating mutations tested in this study. **(c)** Mutation frequency, as measured by the frequency of acquired rifampin resistance ($Rif^R$ mutation frequency, MF) upon expression of AID variants in *E. coli*. Individual data points (n = 2-6) are shown with mean and standard deviation noted, with individual values provided in Source Data Fig. 1. AID-E58A, catalytically inactive control. **(d)** Editing was evaluated in a HEK293T cell line containing *d2gfp*. The presence of *d2gfp*-targeting sgRNA can introduce indels or mutations to generate $GFP^{off}$ cells, which can be tracked by flow cytometry. Shown are representative flow cytometry histograms from transfection of select constructs. Individual data points (n = 3) values are given in Source Data Fig. 1. **(e)** Correlation plot of intrinsic DNA deaminase activity as measured by $Rif^R$ in (**c**) vs. BE activity with average and standard deviation from n≥3 replicates shown.

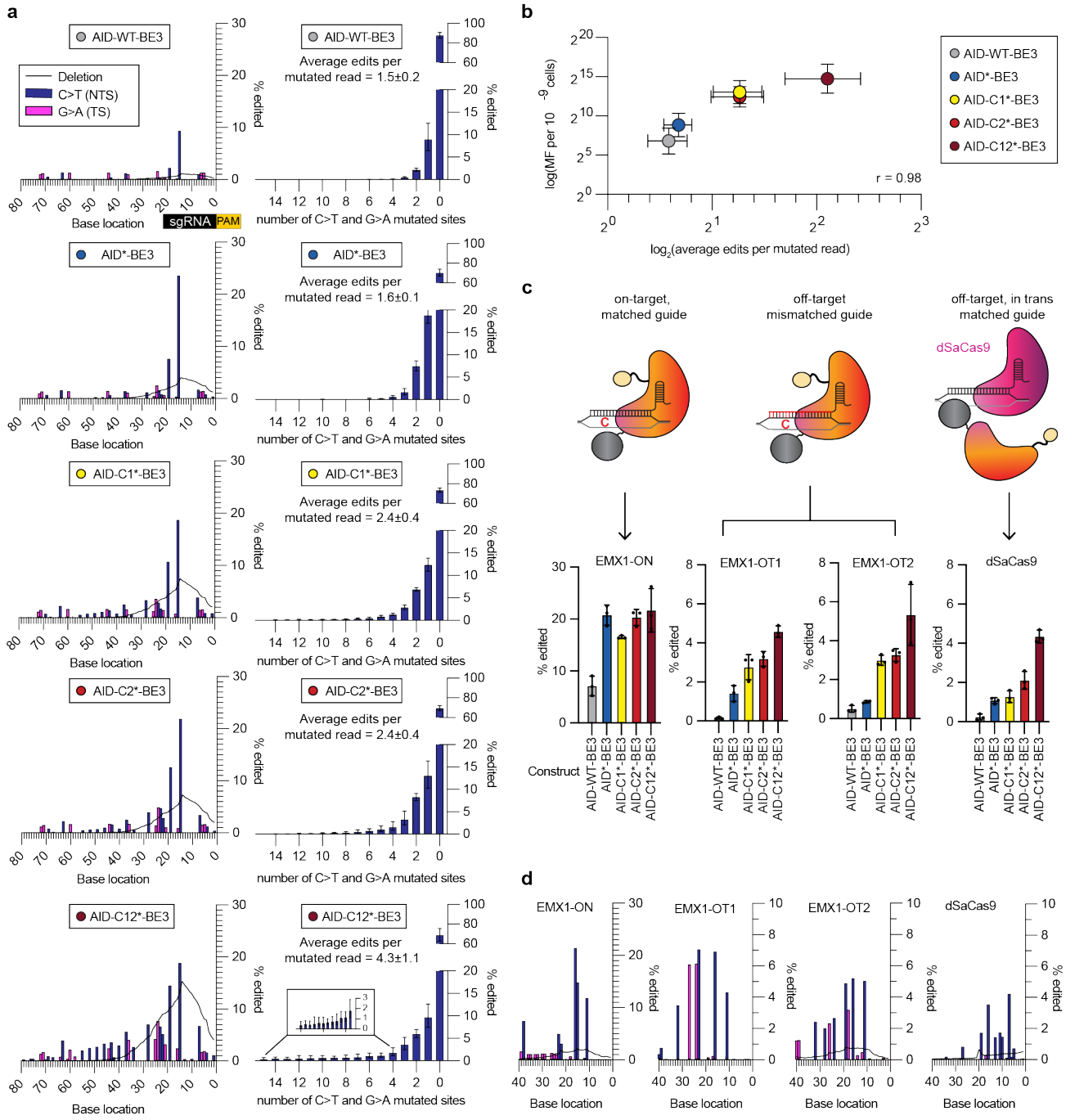



**Figure 2: Hyperactive AID and Cas9 cooperativity establish hBE mutational footprints. (a) Left –** Editing footprints of hBEs across the *d2gfp* locus for each condition. The sgRNA protospacer is located at bases 0-20, with the target C of the Q158 codon at position 16. The PAM is not shown (bases -1 to -3). Data are averaged across three replicates with individual values provided in Source Data Fig. 2. **Right –** Number of C>T and G>A mutations per mutated read for each condition (mean and standard deviation, n=3) with individual values provided in Source Data Fig. 2. **(b)** Correlation plot of intrinsic DNA deaminase activity as measured by Rif$^R$ vs. average number of C>T and G>A mutations per mutated read. **(c)** Shown are deep sequencing C>T conversion efficiency of the target cytosine for three different Cas engagement modes: at the on-target site (EMX1-ON), two sgRNA-dependent off-target sites (EMX1-OT1 and EMX1-OT2) and one sgRNA-independent off-target site (dSaCas9) **(d)** Editing footprints for AID-C12*-BE3 across loci for data in (**c**).

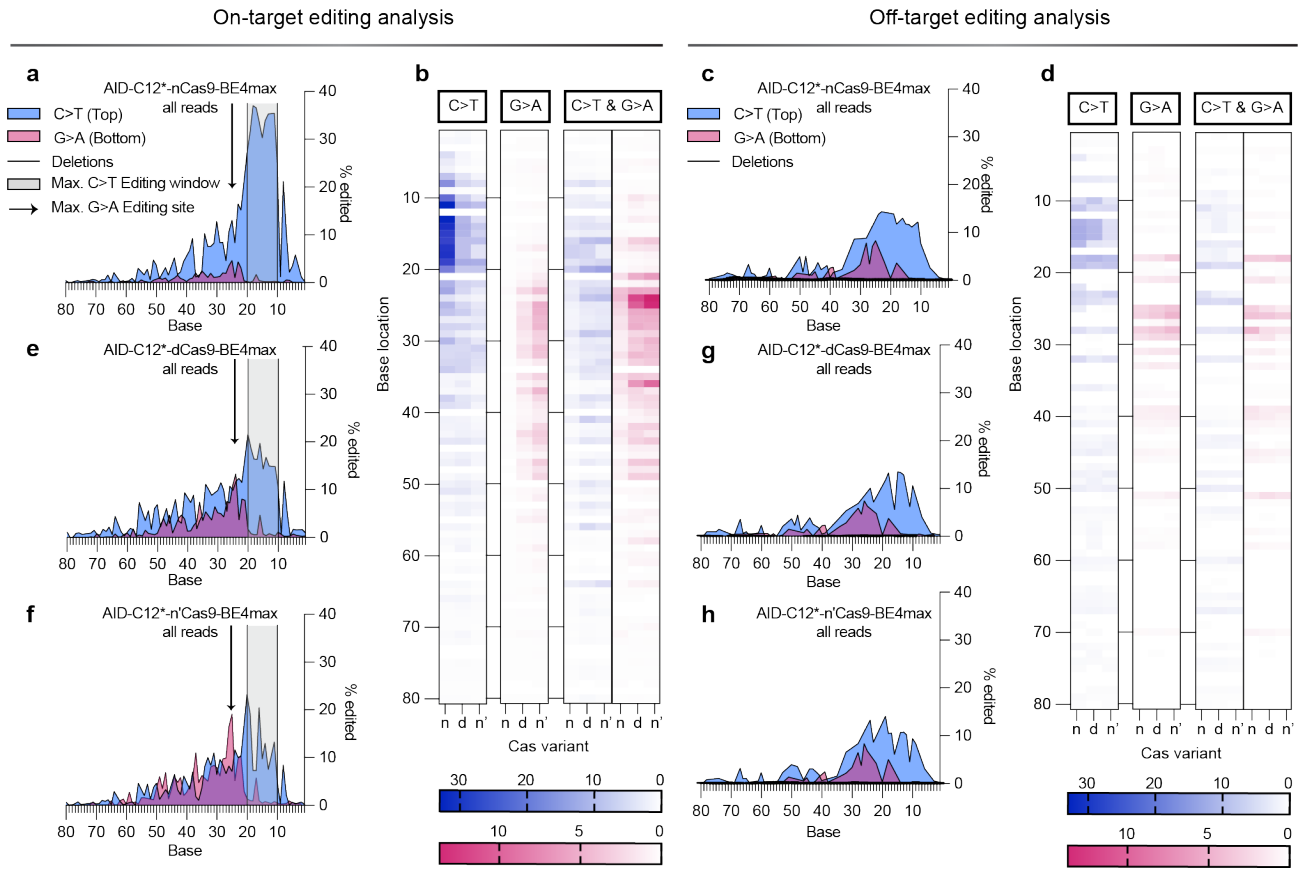



**Figure 3: Perturbing Cas9 cleavage activity skews BE outcomes.** Shown are the averaged editing footprints and heat maps from strand specific analysis of on-target or off-target sites with individual values provided in Source Data Fig. 3. For AID-C12*-BE4max, the editing frequency is averaged across seven loci in (**a**), presented with the sgRNA protospacer located at bases 0-20, while PAM is not shown (-1 to -3). The maximal C>T and G>A editing windows are noted with shading and an arrow, respectively. The heat maps of strand-specific analysis showing C>T only, G>A only and simultaneous C>T/G>A mutations are provided in (**b**) in the corresponding left columns (n, nCas9). A similar editing footprint averaged across four off-target sites is provided in (**c**) with the associated heat maps in (**d**) in the left columns (n, nCas9). Data for individual off-target sites are given in Source Data Fig. 3. The analogous on-target editing footprints for AID-C12*-dCas9-BE4max and AID-C12*-n'Cas9-BE4max are shown in (**e**) and (**f**) respectively, with off-target editing footprints in (**g**) and (**h**). The heatmaps are similarly shown in (**b**) and (**d**), with the dCas9 construct (d) in the middle columns and the n'Cas9 (n') construct at the right ones.

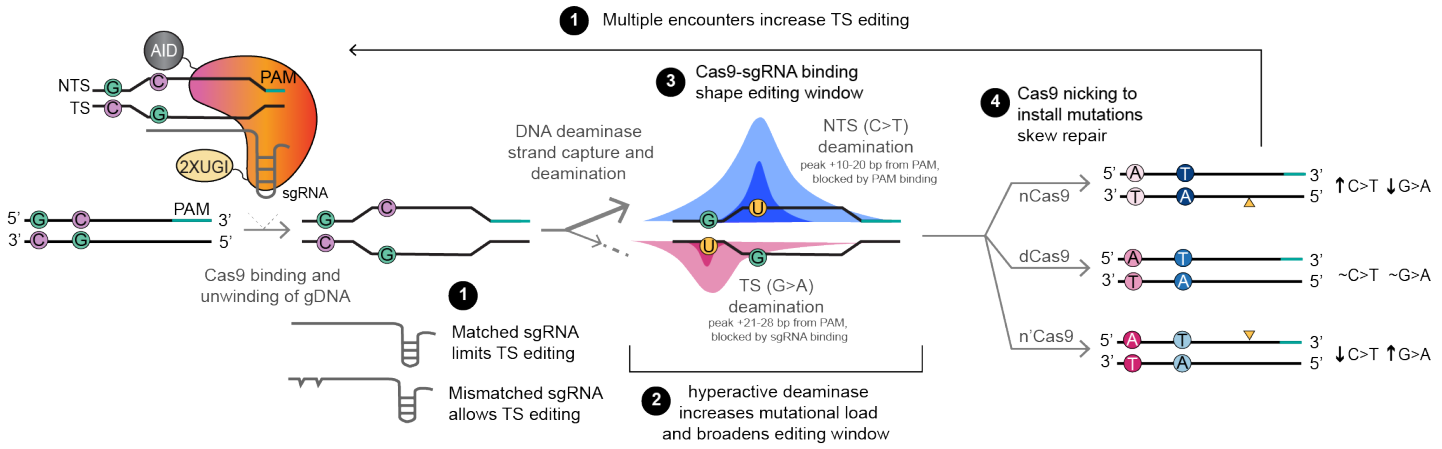



**Figure 4: Cooperativity between Cas9 and AID shapes BE outcomes.** In the mechanistic model, Cas9 and the DNA deaminase work together to determine BE outcomes in multiple manners. **(1)** During an initial hBE encounter, the nature of the R-loop determines whether AID binds the NTS or TS. Matched sgRNA binding promotes NTS editing, while mismatches, which occur at either off-target sites or during second encounters at on-target sites, promote increased TS editing. **(2)** Once engaged with a strand, AID can act processively to introduce multiple C>U mutations, with the number of edits and the size of the editing window depending on the intrinsic deaminase activity. **(3)** Cas9 binding defines the borders of editing, with NTS edits peaking inside the protospacer and TS edits peaking outside the protospacer. **(4)** Finally, the presence of a nick on one strand can increase editing at the other strand. At each step in the mechanism, activity is thus determined by the cooperative actions of the two DNA modifying proteins, Cas9 and AID, acting *in cis*.

## METHODS

*Design and cloning of hyperactive DNA deaminase constructs*

For bacterial studies with AID enzymes, the pET41 expression plasmids containing AID-WT, AID-E58A, or AID* and an N-terminal maltose binding protein tag (MBP), which have been previously described,[31] were used. For hyperactive AID constructs, AID inserts with C1* (D118A, R119G, K120R, A121R), C2* (K10E, T82I, E156G) and C12* (D118A, R119G, K120R, A121R, K10E, T82I, E156G) mutations were obtained as gBlocks (IDT) and cloned into the AID* containing parent construct. The R119G and T82I point mutations were added to the parent AID* construct using the Q5 Site-Directed Mutagenesis Kit (NEB).

For BE3 base editing constructs, the various AID constructs from pET41 constructs were cloned into the scaffold of pCMV_BE3 (Addgene #73021), which contains rat APOBEC1. To this end, the parent pCMV_BE3 plasmid was digested with NotI and XmaI to remove rat APOBEC1. The AID coding sequences were amplified from their respective pET41 constructs, adding flanking regions of overlap with the pCMV_BE3 plasmid backbone. The final plasmids were then constructed using Gibson Assembly Master Mix (NEB), merging the amplified gene fragments with the NotI/XmaI digested parent vector.

For the AID-C12*-nCas9-BE4max construct, AID'-BE4max (Addgene #174696) was used as a scaffold. This parent plasmid contained two point mutations, one in the linker region between Cas9 and UGI and the second one in the first UGI (E11K) protomer. These mutations were corrected here by transferring the NotI/SacI digestion fragment from pCMV-BE4max (Addgene #112093) into AID'-BE4max (Addgene #174696). After generation of this AID-C12*-nCas9-BE4max construct, the point mutations to generate AID-C12*-dCas9-BE4max and AID-C12*-n'Cas9-BE4max were generated using the Q5 Site-Directed Mutagenesis Kit (NEB).

*Bacterial DNA deaminase rifampin mutagenesis assay*

The previously reported rifampin mutagenesis assay[24,31] was used to measure the mutation frequency of AID variants. Plasmids encoding the AID variants (Kanamycin resistant) were transformed into BL21(DE3) *E. coli* harboring a plasmid encoding uracil DNA glycosylase inhibitor (UGI) (chloramphenicol resistant). Overnight cultures were grown in LB with kanamycin (30 ng/mL) and chloramphenicol (25 ng/mL) from single colonies and diluted to an $OD_{600}$ of 0.2. Cells were then grown for 1 hour at 37 °C until they reached $OD_{600}$ of ~0.6 and deaminase expression was induced with 1mM isopropyl β-D-1-thiogalactopyranoside (IPTG). After 4 hours of additional growth, serial dilutions were plated on Luria Bertani (LB) agar plates containing rifampin (100 μg/mL) and plasmid-selective antibiotics (kanamycin and chloramphenicol). The mutation frequencies were then calculated by the ratio of rifampin-resistant ($Rif^R$) colonies relative to the total colony forming units.

*Cell culture*

HEK293T-d2GFP cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) media (Mediatech) supplemented with 10% calf serum (CS) and 1% penicillin/streptomycin at 37 °C with 5% CO2.

*Base editing assay using d2GFP inactivation in HEK293T cells*

The base editing assay was modelled on a previously published protocol.[24] Briefly, HEK293T-d2GFP cells were seeded on 24-well plates and transfected at approximately 60% confluency. 660 ng of the hBE constructs and 330 ng of LRcherry2.1 sgRNA expression plasmids were transfected using 1.5 $\mu$L of Lipofectamine 2000 CD (Invitrogen). Negative control samples include LRcherry2.1T plasmid lacking a protospacer (labeled as no sgRNA samples). The *d2gfp*-targeting sgRNA can support the introduction of indels in *d2gfp* with Cas9 or a Q158X nonsense mutation with base editing, along with other potential edits. Transfected cells were harvested at day 3 after transfection, ensuring single-cell suspension. Genomic DNA was collected from cells using the DNeasy Blood & Tissue Kit (Qiagen) and the *d2GFP* locus was amplified and then deep sequenced as described on the "DNA library preparation" section below. Representative experiments were repeated independently three times and the results were reproducible.

*Base editing of various genomic loci*

The base editing assay for diverse genomic loci was based of a previously published protocol.[24] Briefly, the transfection protocol was performed as above, but with different sgRNAs targeting various loci instead of the *d2gfp*-targeting sgRNA. Transfected cells were harvested at day 3 after transfection, ensuring single-cell suspension. Genomic DNA was collected using the DNeasy Blood & Tissue Kit (Qiagen) and each locus was amplified and then deep sequenced as described on the "DNA library preparation" section below. Representative experiments were repeated independently three times and the results were reproducible.

*R-loop assay*

The R-loop assay used in this study was based of a previously published protocol.[24] Briefly, HEK293T-d2GFP cells were seeded on 24-well plates and transfected at ~60% confluency. 400 ng of hBE constructs, 200 ng of *EMX1*-targeting LRcherry2.1 sgRNA plasmid, and 400 ng of dSaCas9 expression plasmid were co-transfected using 1.5 $\mu$L of Lipofectamine 2000 CD (Invitrogen) per well. Transfected cells were harvested at day 3 after transfection, ensuring single-cell suspension. Genomic DNA was collected from cells using the DNeasy Blood & Tissue Kit (Qiagen) and both the *EMX1, EMX1-OT1, EMX1-OT2* and dSaCas9-targeted locus (Chr 9: 21036-21332) were amplified and then deep sequenced as described on the "DNA library preparation" section below. Representative experiments were repeated independently three times and the results were reproducible.

*DNA library preparation*

Genomic DNA was collected from cells using the DNeasy Blood & Tissue Kit (Qiagen). The target loci of interest were PCR-amplified from 100 ng genomic DNA using KAPA HiFi HotStart Uracil+ Ready Mix (Kapa Biosystems) or Phusion High-Fidelity DNA Polymerase (New England Biolabs, NEB). PCR products were purified via QIAquick PCR Purification Kit (Qiagen). Indexed DNA libraries were prepared using the NEBNext Ultra II DNA Library Prep Kit (Illumina) with the following specifications. After adapter ligation and 4 cycles of PCR enrichment, indexed amplicon concentration was quantified by Qubit dsDNA HS Assay Kit (ThermoFisher). Indexed PCR amplicons were pooled together in an equimolar ratio for paired-end sequencing by MiSeq (Illumina) with the 300-cycle MiSeq Reagent Nano Kit v2 (Illumina) or 300-cycle MiSeq Reagent Micro Kit v2 (Illumina).

*Statistics*

Statistics were performed using GraphPad Prism. Comparisons of > 2 groups was performed using one-way ANOVA adjusted for Tukey's multiple comparisons. Information on statistical tests performed, exact values of n, and how significance was defined is available in the figure legends or provided as source data. For cell culture experiments, n is defined as cell cultures treated with separately prepared transfection solutions in different days.

*Flow cytometry*

For GFP editing analysis, HEK293T-d2GFP cells were harvested and 50,000 total cell events were gated with FSC/SSC gate on a Guava Easycyte 10HT instrument (Millipore). A second gate for high mCherry transfection efficiency was applied. The $GFP^{off}$ population gating was performed using the nadir between the two peaks, with consistent gates for all comparison experiments. Flow cytometry analysis was performed using FlowJo Software Version 10.7.1 (FloJo, LCC).

*Next-Generation Sequencing Analysis*

Raw reads were automatically demultiplexed by MiSeq Reporter. Demultiplexed read qualities were evaluated by FastQC v0.11.9 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality sequence (Phred quality score <28) and adapters were trimmed via Trim Galore v0.6.5 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) prior to analysis with CRISPResso2.[32] Sequencing yielded ~1,000-50,000 aligned reads per sample.

*Data Processing Scripts*

The following R scripts use raw .txt files (Alleles_frequency_table.txt) displaying all reads (all DNA sequences) produced from CRISPResso 2.0 alignment of Next Generation Sequencing (NGS) data to a reference sequence. Each row of the .txt file represents a different read sequence, and the abundance of this sequence is given as a number of reads. In the following descriptions, DNA amount and the number of reads are used synonymously at times. This is because the abundance of a read sequence is representative of the total population of DNA.

Before script-specific data tidying, there is common data cleanup that occurs:
Reads with insertions and deletions (indels) are discarded from the analysis, given the interest in C>T and G>A editing. Further, analysis is focused on the window of interest by stripping read sequences to 80 base positions 5' of the PAM sequence. In data inspection of the CRISPResso output, some reads were misaligned to the reference sequence. To minimized misalignment, all reads were aligned with the reference sequence in the script and reads that contain all permutations of mutations were eliminated as a marker of frameshifts.

**1Generalized_BarGraphMutatedSites**

This R script produces data that reflects the overall mutagenic activity of the base editor being evaluated.

First, the common analysis described above is followed. The *n* number of intended mutations (for e.g., C>T mutations for cytosine base editors) on the target strand (TS) and nontarget strand (NTS) is then counted. Since reads represent the NTS, the script looks for both C>T (C>T mutations on the NTS) and G>A (C>T mutations on the TS). *n* becomes an identifier for each read sequence. The number of reads that has the same *n* are then collated. At this point, the data are in the form of "*m* number of reads have *n* mutations". The percentage of reads with *n* number of mutations is then determined by dividing *m* by the total number of reads in the sample (this number excludes indels and misaligned reads given the preceding common data cleanup that occurs). The data are then extracted in the form of "X% of DNA have *n* number of mutations created by the base editor".

**2Processivity_Average**

This R script produces data that shows how editing activity changes by base location.

First, the common analysis described above is followed. To a working .txt file whose rows represent different read sequences, 80 empty columns are added. Each added column represents a different base position 5' of the PAM sequence. Each read sequence is then parsed for the mutation of interest, and the base position the edit occurs in is recorded by filling in a "1" for an edited base in the respective column. Weights are then added to these recordings according to the number of times the same kind of read sequence is detected. This is achieved by multiplying the numbers in each row of these 80 columns with the number of reads per sequence existing in the sample pool. Next, each column is summed, and the sums for each column are divided by the total number of reads (this number excludes indels and misaligned reads). This data reads as "on *x* base position, *y*% of reads were edited". As a target mutation cannot occur at every base position (a C>T will only occur on cytosines), the analysis is repeated on multiple loci that have different target sequences and averaged to construct a sequence of 80 base positions with an editable base at almost every position. After repeating this analysis, the <u>nonzero</u> *y*%, corresponding to an editable base across several or all loci, are averaged for all loci containing this editable base. After averaging, the data are extracted in the form of "*y*% (averaged) of DNA would be edited at *x* base position 5' of the PAM sequence if every* base could be edited".

*Across all loci, every base cannot be an editable base despite averaging, as a cytosine might not exist in a certain position in all of the loci analyzed

This analysis can be performed for multiple mutations independently by changing the mutation of interest, and instead of carrying out this analysis on all reads produced by the CRISPResso output, it is possible to subset to a group of reads (such as reads with mutation only on the TS, only on the NTS, reads with the target mutation on both strands etc.). To subset, the reads are parsed for mutations on both the TS and NTS, tagged with an identifier for either one or both strands, and then filtered for this identifier before proceeding with the addition of the 80 new columns and subsequent analysis.

**3AllTypes_Processivity**

This R script produces data that shows how editing efficiency changes as the position of a base of interest increases 5' away from the PAM.

The script is different from **2Processivity Average** as it only runs on one locus, and before analysis, deletions are not discarded. Otherwise, the common analysis described above is followed. As with **2Processivity Average,** 80 empty columns are added to the .txt file and weights are added similarly. Next, each column is summed, and the sums for each column are divided by the total number of reads (this number excludes indels and misaligned reads). This data reads as "on *x* base position, *y*% of reads were edited". Again, as the target mutation cannot occur at every base position (a C>T will only occur on cytosines), for positions that do not have an editable base, *y*% is changed from 0% to N/A. Next, the analysis is repeated by adding a new set of 80 columns, this time looking for a "-" instead of the mutation of interest. This records the location of deletions, which can then be used to similarly calculate the *w*% of reads with deletions at *x* base position. The data are extracted in the form of "*y*% and *w*% of the sample DNA is edited and has a deletion, respectively, at *x* base position 5' of the PAM sequence of *z* locus", where *z* is the locus being analyzed. Like **2Processivity Average,** this analysis can be carried out for multiple mutations independently and subsets of reads. The subsetting would occur before adding and recording the mutations or deletions in the 80 new columns.

## DATA AVAILABILITY

Novel plasmids generated in this study will be deposited to Addgene. Next-Generation Sequencing (NGS) data raw fastQ files were deposited to the National Center for Biotechnology Information (NCBI) - BioProject: PRJNA891058. Unprocessed NGS data from CRISPResso2 is available in the Source Data Files. Author's original code and the NGS data from CRISPResso2 processed with the code is available in the supplementary information and Source Data Files respectively. Further information and requests for reagents should be directed to and will be fulfilled by the Lead Contacts, Rahul M. Kohli (rkohli@pennmedicine.upenn.edu) and Junwei Shi (jushi@pennmedicine.upenn.edu).

## CODE AVAILABILITY

Original Codes to analyze NGS data are supplied in .txt files for use in Rstudio.

## REFERENCES


1. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824-844 (2020).

2. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424 (2016).

3. Zong, Y., Song, Q., Li, C., Jin, S., Zhang, D., Wang, Y., Qiu, J. L. & Gao, C. Efficient C-to-T base editing in plants using a fusion of nCas9 and human APOBEC3A. *Nat. Biotechnol.* (2018).

4. Ma, Y., Zhang, J., Yin, W., Zhang, Z., Song, Y. & Chang, X. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat. Methods* **13**, 1029-1035 (2016).

5. Nishida, K., Arazoe, T., Yachie, N., Banno, S., Kakimoto, M., Tabata, M., Mochizuki, M., Miyabe, A., Araki, M., Hara, K. Y., Shimatani, Z. & Kondo, A. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, 10.1126/science.aaf8729. Epub 2016 Aug 4 (2016).

6. Hess, G. T., Fresard, L., Han, K., Lee, C. H., Li, A., Cimprich, K. A., Montgomery, S. B. & Bassik, M. C. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* **13**, 1036-1042 (2016).

7. Ren, B., Yan, F., Kuang, Y., Li, N., Zhang, D., Zhou, X., Lin, H. & Zhou, H. Improved Base Editor for Efficiently Inducing Genetic Variations in Rice with CRISPR/Cas9-Guided Hyperactive hAID Mutant. *Mol. Plant.* (2018).

8. Xiong, X., Li, Z., Liang, J., Liu, K., Li, C. & Li, J. F. A cytosine base editor toolkit with varying activity windows and target scopes for versatile gene manipulation in plants. *Nucleic Acids Res.* **50**, 3565-3580 (2022).

9. Kim, Y. B., Komor, A. C., Levy, J. M., Packer, M. S., Zhao, K. T. & Liu, D. R. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat. Biotechnol.* **35**, 371-376 (2017).

10. Koblan, L. W., Doman, J. L., Wilson, C., Levy, J. M., Tay, T., Newby, G. A., Maianti, J. P., Raguram, A. & Liu, D. R. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843-846 (2018).

11. Zafra, M. P., Schatoff, E. M., Katti, A., Foronda, M., Breinig, M., Schweitzer, A. Y., Simon, A., Han, T., Goswami, S., Montgomery, E., Thibado, J., Kastenhuber, E. R., Sanchez-Rivera, F. J., Shi, J., Vakoc, C. R., Lowe, S. W., Tschaharganeh, D. F. & Dow, L. E. Optimized base editors enable efficient editing in cells, organoids and mice. *Nat. Biotechnol.* **36**, 888-893 (2018).

12. Liu, Huang, Dai, Liu, Fan, Cheng, Zhao, Yeap & Meng. Intrinsic Nucleotide Preference of Diversifying Base Editors Guides Antibody Ex Vivo Affinity Maturation. *Cell reports (Cambridge)* **25**, 884-892.e3 (2018).

13. Gehrke, J. M., Cervantes, O., Clement, M. K., Wu, Y., Zeng, J., Bauer, D. E., Pinello, L. & Joung, J. K. An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. *Nat. Biotechnol.* **36**, 977-982 (2018).

14. Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I. & Liu, D. R. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **551**, 464-471 (2017).

15. Huang, T. P., Zhao, K. T., Miller, S. M., Gaudelli, N. M., Oakes, B. L., Fellmann, C., Savage, D. F. & Liu, D. R. Circularly permuted and PAM-modified Cas9 variants broaden the targeting scope of base editors. *Nat. Biotechnol.* **37**, 626-631 (2019).

16. Kleinstiver, B. P., Whittaker, M. N., Christie, K. A. & Walton, R. T. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290-296 (2020).

17. Conticello, S. G., Langlois, M. A., Yang, Z. & Neuberger, M. S. DNA deamination in immunity: AID in the context of its APOBEC relatives. *Adv. Immunol.* **94**, 37-73 (2007).

18. Di Noia, J. M. & Neuberger, M. S. Molecular Mechanisms of Antibody Somatic Hypermutation. *Annual review of biochemistry* **76**, 1-22 (2007).

19. Ghorbani, A., Quinlan, E. M. & Larijani, M. Evolutionary Comparative Analyses of DNA-Editing Enzymes of the Immune System: From 5-Dimensional Description of Protein Structures to Immunological Insights and Applications to Protein Engineering. *Front. Immunol.* **12**, 642343 (2021).

20. Wang, M., Yang, Z., Rada, C. & Neuberger, M. S. AID upmutants isolated using a high-throughput screen highlight the immunity/cancer balance limiting DNA deaminase activity. *Nat. Struct. Mol. Biol.* **16**, 769-776 (2009).

21. Gajula, K. S., Huwe, P. J., Mo, C. Y., Crawford, D. J., Stivers, J. T., Radhakrishnan, R. & Kohli, R. M. High-throughput mutagenesis reveals functional determinants for DNA targeting by activation-induced deaminase. *Nucleic Acids Res.* **42**, 9964-9975 (2014).

22. Barreto, V., Reina-San-Martin, B., Ramiro, A. R., McBride, K. M. & Nussenzweig, M. C. C-terminal deletion of AID uncouples class switch recombination from somatic hypermutation and gene conversion. *Mol. Cell* **12**, 501-508 (2003).

23. Wang, M., Rada, C. & Neuberger, M. S. A high-throughput assay for DNA deaminases. *Methods Mol. Biol.* **718**, 171-184 (2011).

24. Berríos, K. N., Evitt, N. H., DeWeerd, R. A., Ren, D., Luo, M., Barka, A., Wang, T., Bartman, C. R., Lan, Y., Green, A. M., Shi, J. & Kohli, R. M. Controllable genome editing with split-engineered base editors. *Nat. Chem. Biol.* **17**, 1262-1270 (2021).

25. 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.* (2020).

26. Zhang, X., Zhu, B., Chen, L., Xie, L., Yu, W., Wang, Y., Li, L., Yin, S., Yang, L., Hu, H., Han, H., Li, Y., Wang, L., Chen, G., Ma, X., Geng, H., Huang, W., Pang, X., Yang, Z., Wu, Y., Siwko, S., Kurita, R.,

Nakamura, Y., Yang, L., Liu, M. & Li, D. Dual base editor catalyzes both cytosine and adenine base conversions in human cells. *Nat. Biotechnol.* **38**, 856-860 (2020).

27. Levy, J. M., Yeh, W. H., Pendse, N., Davis, J. R., Hennessey, E., Butcher, R., Koblan, L. W., Comander, J., Liu, Q. & Liu, D. R. Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nat. Biomed. Eng.* **4**, 97-110 (2020).

28. Grünewald, J., Zhou, R., Lareau, C. A., Garcia, S. P., Iyer, S., Miller, B. R., Langner, L. M., Hsu, J. Y., Aryee, M. J. & Joung, J. K. A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing. *Nat. Biotechnol.* **38**, 861-864 (2020).

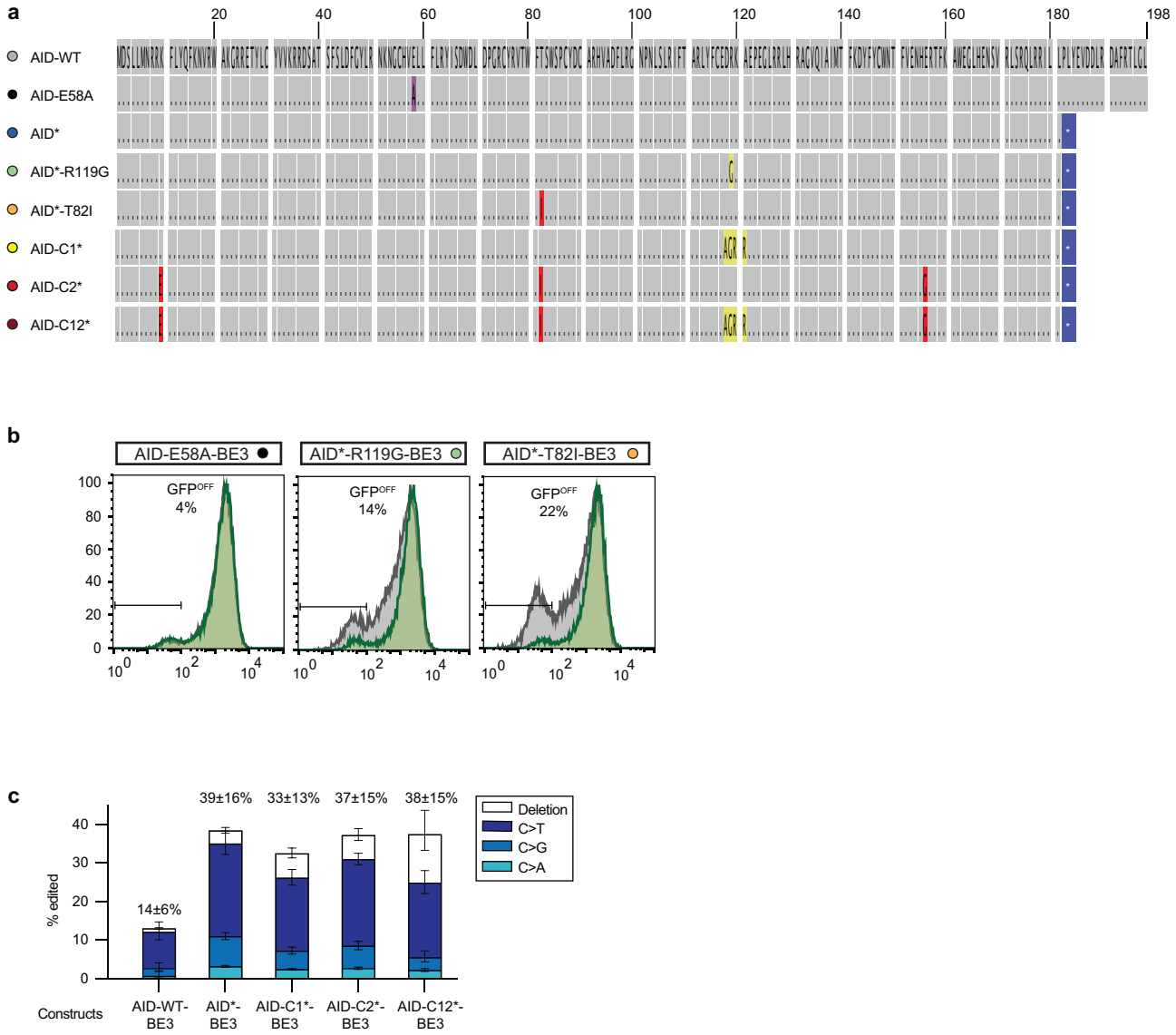
29. Li, C., Zhang, R., Meng, X., Chen, S., Zong, Y., Lu, C., Qiu, J. L., Chen, Y. H., Li, J. & Gao, C. Targeted, random mutagenesis of plant genes with dual cytosine and adenine base editors. *Nat. Biotechnol.* **38**, 875-882 (2020).

30. Qiao, Q., Wang, L., Meng, F. L., Hwang, J. K., Alt, F. W. & Wu, H. AID Recognizes Structured DNA for Class Switch Recombination. *Mol. Cell* **67**, 361-373.e4 (2017).

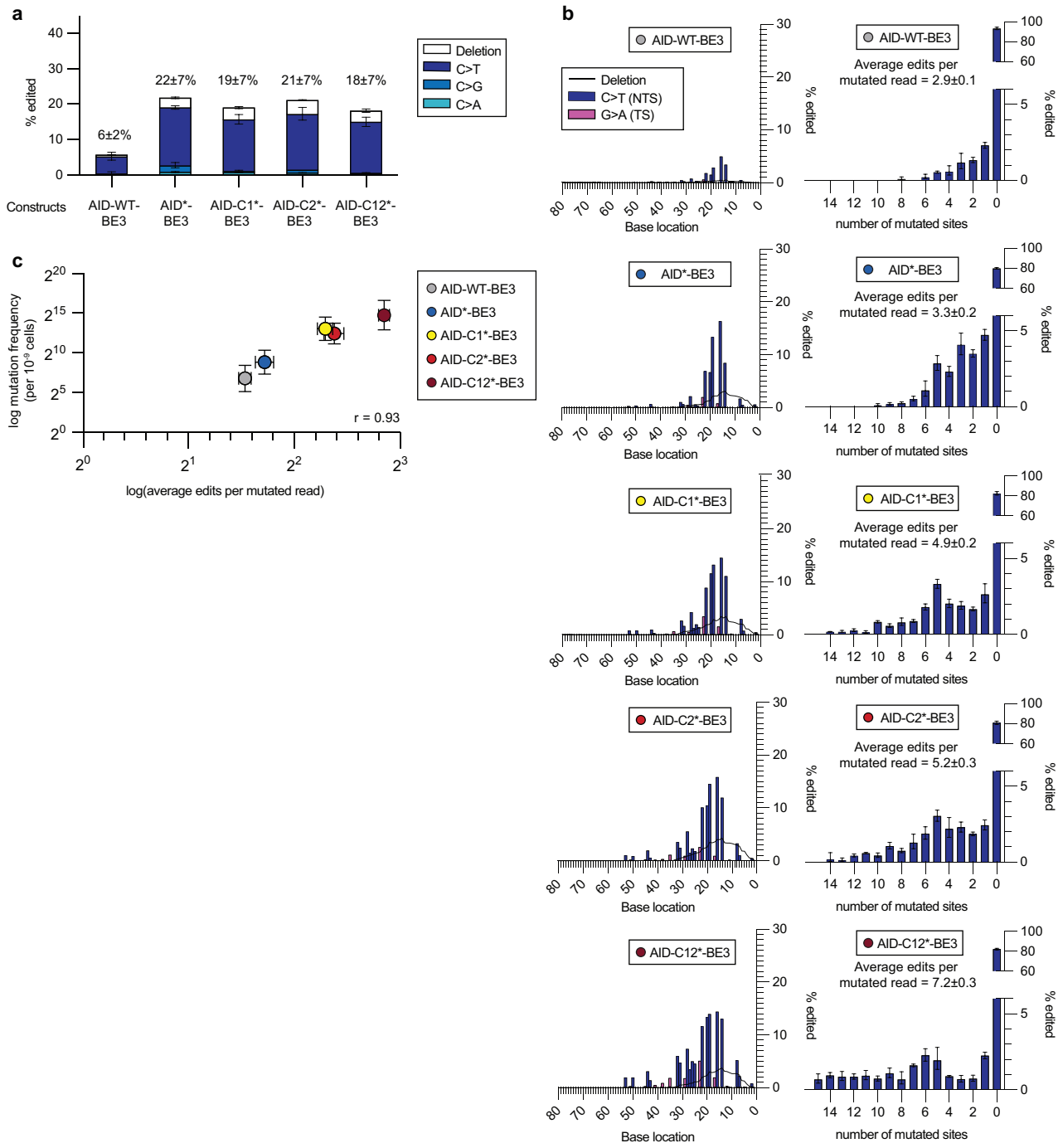
31. Kohli, R. M., Abrams, S. R., Gajula, K. S., Maul, R. W., Gearhart, P. J. & Stivers, J. T. A portable hotspot recognition loop transfers sequence preferences from APOBEC family members to activation-induced cytidine deaminase. *J. Biol. Chem.* **284**, 22898-22904 (2009).

32. Clement, K., Rees, H., Canver, M. C., Gehrke, J. M., Farouni, R., Hsu, J. Y., Cole, M. A., Liu, D. R., Joung, J. K., Bauer, D. E. & Pinello, L. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224-226 (2019).

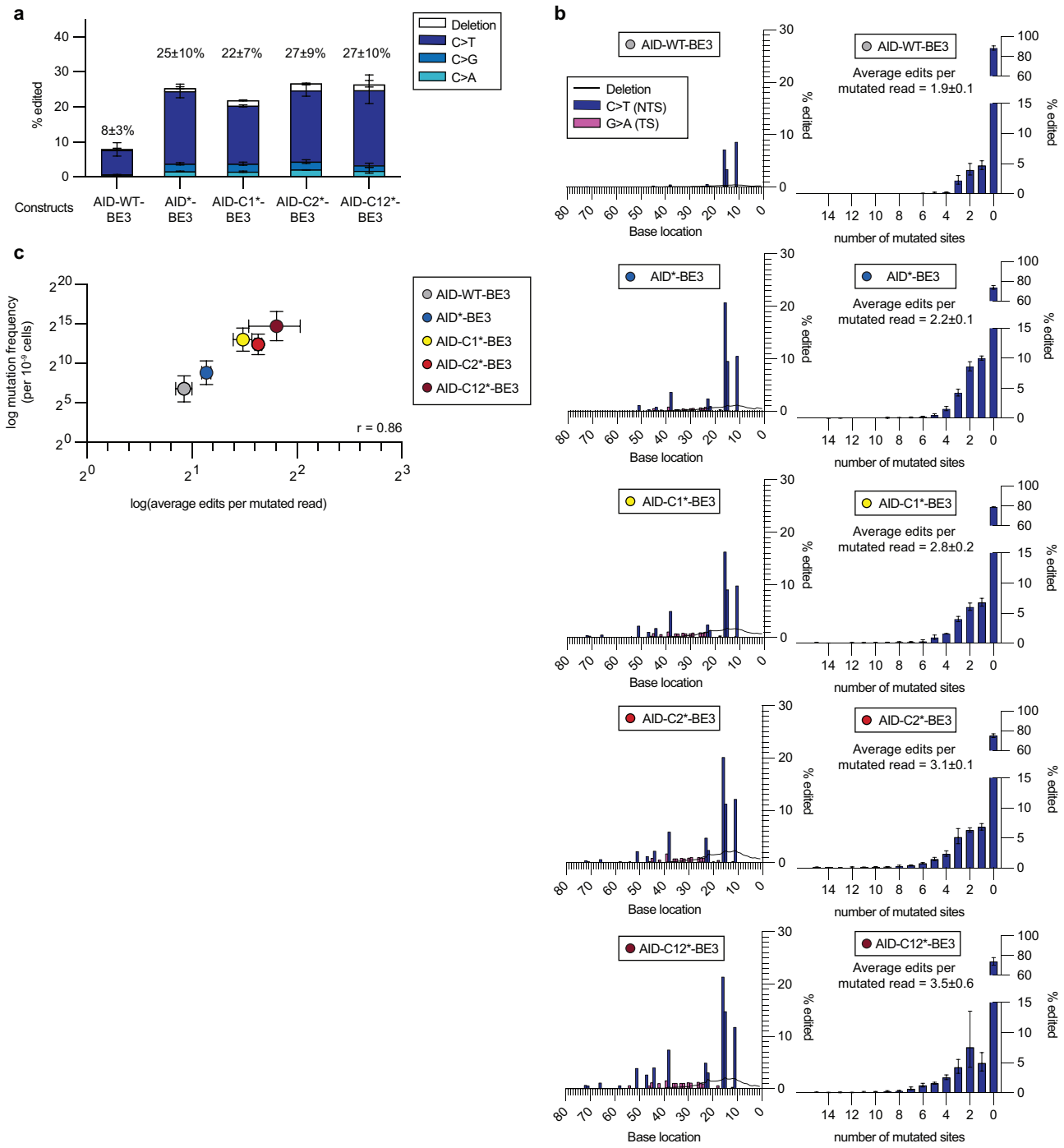
# EXTENDED DATA FIGURES



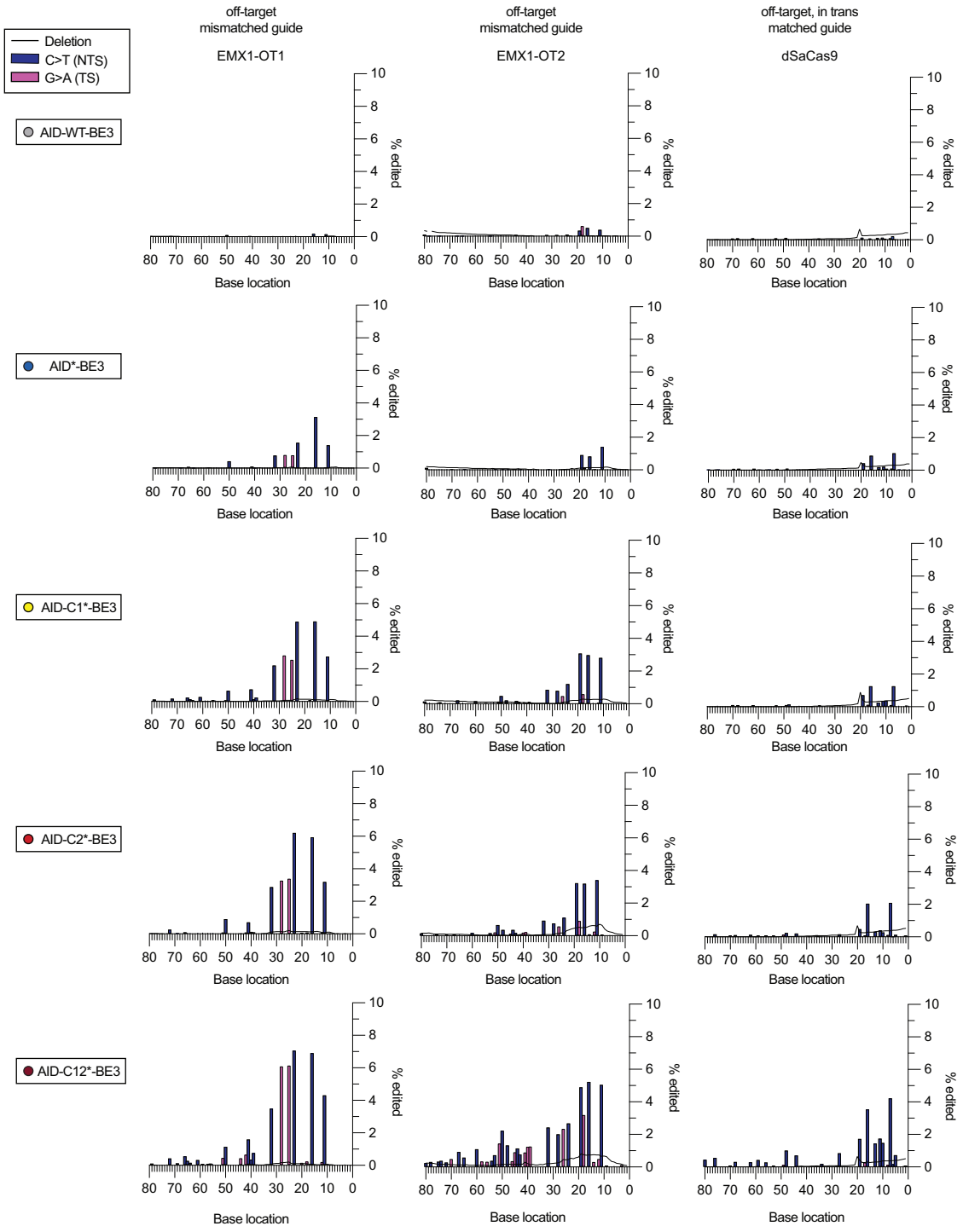

**Extended Data Figure 1: AID variants and hBEs flow cytometry plots.** **(a)** Schematic of AID protein sequence detailing the mutations of each variant deaminase tested. **(b)** Representative flow cytometry histograms associated with transfection of hBEs in HEKT293T-d2GFP cells for constructs not shown in **Fig. 1**. **(c)** Deep sequencing shows conversion efficiency at the target C to T, G, A and deletions (mean and standard deviation, n=3) with individual values provided in Source Data Extended Data Fig. 1.



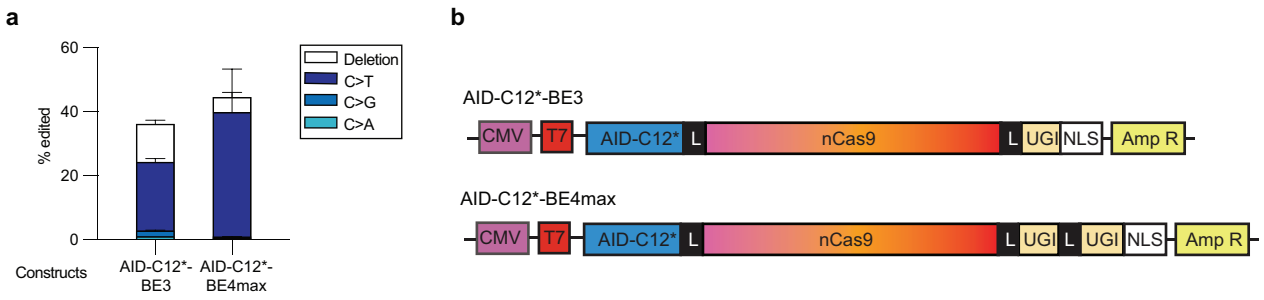

**Extended Data Figure 2: hBE activity at the *TDG* locus. (a)** Deep sequencing results with BE constructs demonstrating conversion efficiency at the target C to T, G, A and deletions (mean and standard deviation, n=3). **(b) Left -** Editing footprints of hBEs across the *TDG* locus for each condition. The sgRNA protospacer is located at bases 0-20. The PAM is not shown (bases -1 to -3). Data are averaged across three replicates. **Right –** Number of C>T and G>A mutations per mutated read for each condition (mean and standard deviation, n=3). **(c)** Correlation plot of intrinsic DNA deaminase activity as measured by $Rif^R$ vs. average number of C>T and G>A mutations per mutated read.




**Extended Data Figure 3: hBE activity at the *EMX1* locus. (a)** Deep sequencing results with BE constructs demonstrating conversion efficiency at the target C to T, G, A and deletions (mean and standard deviation, n=3). **(b) Left -** Editing footprints of hBEs across the *EMX1* locus for each condition. The sgRNA protospacer is located at bases 0-20. The PAM is not shown (bases -1 to -3). Data are averaged across three replicates. **Right –** Number of C>T and G>A mutations per mutated read for each condition (mean and standard deviation, n=3). **(c)** Correlation plot of intrinsic DNA deaminase activity as measured by $Rif^R$ vs average number of C>T and G>A mutations per mutated read.




**Extended Data Figure 4: Off-target hBE activity.** Editing footprints of hBEs are shown across two sgRNA-dependent off-target sites (*EMX1-OT1* and *EMX-OT2)* and one sgRNA-independent off-target site (dSaCas9 binding site) generated by the R-loop assay. The sgRNA protospacer is located at bases 0-20. The PAM is not shown (bases -1 to -3). Data are averaged across three replicates.




**Extended Data Figure 5: BE3 and BE4max comparison. (a)** Deep sequencing comparing target cytosine (C) conversion efficiency to T, G, A and deletions by BE3 and BE4max (mean and standard deviation, n=3) with individual values in Source Data Extended Data Fig. 5. **(b)** Construct schematics for AID-C12*-BE3 and AID-C12*-BE4max.

## Supplementary Table 1 - Sequencing Sheet

| target gene | sgRNA sequence | Target strand | Forward Primer | Reverse Primer | Size | Amplicon Sequence | Chromosome | Start Position | End Position |
|---|---|---|---|---|---|---|---|---|---|
| d2GFP | CAAGCAGAAGAACGGCATCA | + | CTTCAAGGAGGACGGCAAC | GTGGTCGGCGAGCTG | 187 | ATTTTCAGGGGTCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACCTAGGCTGCTAAACA | N/A | | |
| CDA | ATCCAGAAGGCCGTCTCAGA | – | CTCTCTTCTCAGGGTGCAACATAG | CAACCTGGCTTTCCCACTCAC | 145 | CTCTCTTCTCAGGGTGCAACATAGAAAATGCCTGCTACCCGCTGGGCATCTGTGCTGAACGGACCGCTATCCAGAAGGCCGTCTCAGAAGGGTACAAGGATTTCAGGGCAATTGCTATCGCCAGGTGAGTGGGAAAGCCAGGTTG | 1 | 20604916 | 20605060 |
| TDG_g1 | CCTGCTCAGGAACCAGTGCA | + | CAACAACTGATGGCTGAAGCTC | GCTGGCTGATCCGATGTTGAAC | 148 | CAACAACTGATGGCTGAAGCTCCTAATATGGCAGTTGTGAATGAACAGCAAATGCCAGAAGAAGTTCCAGCCCCAGCTCCTGCTCAGGAACCAGTGCAAGGTAGTTTCACCATGAGAGCTTAGAAAGTTCAACATCGGATCAGCCAGC | 12 | 103976983 | 103977108 |
| TDG_g2 | TGAGGTCCAGCTGAACCATA | + | GATCGTGCCACTACACTCTAGC | GCGTGGTCCTTTCCACCATG | 197 | GATCGTGCCACTACACTCTAGCCTGGGTGACAGAGCGAGACCCTGTCTAAAAAAAAATAAATAACTGAATTTTCATTTTACAGGGAAGTGTTTGTTTATGTCAGGGCTCAGTGAGGTCCAGCTGAACCATATGGATGATCACACTCTACCAGGGAAGTATGGTATTGGATTTACCAACATGGTGGAAAGGACCACGC | 12 | 103982716 | 103982912 |
| IDH2 | GCGTGCCTGCCAATGGTGAT | – | CCCAATGGAACTATCCGGAAC | CTAGGCGAGGAGCTCCAGTC | 196 | CCCAATGGAACTATCCGGAACATCCTGGGGGGGACTGTCTTCCGGGAGCCCATCATCTGCAAAAACATCCCACGCCTAGTCCCTGGCTGGACCAAGCCCATCACCATTGGCAGGCACGCCCATGGCGACCAGGTAGGCCAGGGTGGAGAGGGGATCCACTGACCTGGGCACCCCCCGACTGGAGCTCCTCGCCTAG | 15 | 90088523 | 90088718 |
| EMX1 | GAGTCCGAGCAGAAGAAGAA | + | CCCATCAGGCTCTCAGCTCAGCC | CGATGTCCTCCCCATTGGCCTGC | 246 | CCCATCAGGCTCTCAGCTCAGCCTGAGTGTTGAGGCCCCAGTGGCTGCTCTGGGGGCCTCCTGAGTTTCTCATCTGTGCCCCTCCCTCCCTGGCCCAGGTGAAGGTGTGGTTCCAGAACCGGAGGACAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGAAGCAGGCCAATGGGGAGGACATCG | 2 | 72933689 | 72933934 |
| FANCF | GGAATCCCTTCTGCAGCACC | + | GACCAATAGCATTGCAGAGAGGCG | GTGCTGACGTAGGTAGTGCTTGA | 181 | GACCAATAGCATTGCAGAGAGGCGTATCATTTCGCGGATGTTCCAATCAGTACGCAGAGAGTCGCCGTCTCCAAGGTGAAAGCGGAAGTAGGGCCTTCGCGCACCTCATGGAATCCCTTCTGCAGCACCTGGATCGCTTTTCCGAGCTTCTGGCGGTCTCAAGCACTACCTACGTCAGCAC | 11 | 22625737 | 22625917 |
| EMX1_OFF1 | TTCTTCTTCTGCTTAGACTC | – | CAAGGATGCAGTCTCATGACTTGG | CCATTGAAATCTCACCTGGGCG | 210 | CAAGGATGCAGTCTCATGACTTGGCCTTTGTAGGAAAACACCATTAGAAGAGTAGATGGTTGGGTAGTGGCTCTCTTCTTCTTCTGCTTAGACTCTTGTCTACTATGAATAAAGGGCTCTATTTGCAAAGGCCGTGATGGGTTGAAGCACATTGAGAAAGAGGCTACTGAAACATAAGTTACCTTTCTCGCCCAGGTGAGATTTCAATGG | 15 | 43817434 | 43817643 |
| EMX1_OFF2 | TTCTTCTTCTGCTCAGACGT | – | CTCTCGTCTTCCTGCAGAGGTTC | CAGTCAACAAAGCCAGCCTCATA | 307 | CTCTCGTCTTCCTGCAGAGGTTCTGCCAGTGCCTCAAGAATGTCCTCAGAGTTCTGTCCATTCTTCTTCTGCTCAGACGTTTTGTCTGAAAGTATTTTCCCAGGCCAGAGCTGGACTCTGTATAGAGTACAGAGGAAGGGCCAAGGTATGACTTTACTCCCACCACTTATATCTGAGAGCACAGGCACAGAGGGCAGTTTGTCACAAATCTGAGTGTGCCCGAGCTACATTCATGAATAATAAATAGGCTTCATATTGCCTCACATTCCCATTCCCATCCCAAGTATGAGGCTGGCTTTGTTGACTG | 6 | 9118503 | 9118809 |
| FANCF_OFF1 | GGAACCCCGTCTGCAGCACC | – | CAGGAGCCGGCTTTCTGTGTC | CTCTCAAGTCACCTGGATCGTCC | 222 | CAGGAGCCGGCTTTCTGTGTCAGGAGAGGCCACGCCAGCACTTTCTAAGGAATCGCGGCGCCTGCTGCACTCTCTGAGTATTTTCCGGCAGGGCCTGCGGGCAGTGGCGTCTTAGTCGCCTTAGCACTGGGTGCTTAATCCGGCTCCATCTTTTCTCCACGGAGGGGGCCTGGTGCTGCAGACGGGGTTCCCGGGGTCAGGACGATCCAGGTGACTTGAGAG | 18 | 8707356 | 8707577 |
| FANCF_OFF2 | ACCATCCCTCCTGCAGCACC | + | GCCAGTGAACTAGAGTGGCATG | GCCTCACAACTCTGCCATGTG | 228 | GCCAGTGAACTAGAGTGGCATGCAACCTAGAGAGATACTCACTGGCATGACTGAGAGATGGTTTGTACCATCCCTCCTGCAGCACCAGGCAGCAAAACTCATAGCAACTAAAGTGACTCATTTCTCCTGCTTAAGGAAAAGAAATCAAAGTATAAAGAGGATTTTGTCTTGCATCTTGGATACCATCTCATTCACAATAGGATAAGGCACATGGCAGAGTTGTGAGGC | X | 87100094 | 87100321 |